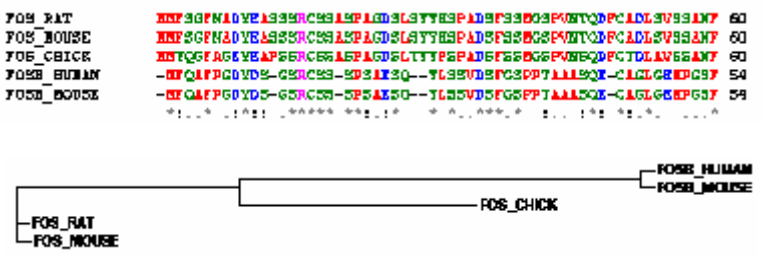
ΕΡΓΑΣΤΗΡΙΑΚΕΣ ΣΗΜΕΙΩΣΕΙΣ

ΕΦΑΡΜΟΣΜΕΝΗΣ

ΒΙΟΜΕΤΡΙΑΣ-ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ



ΗΛΙΑΣ ΖΙΝΤΖΑΡΑΣ, M.Sc., Ph.D.
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ
ΒΙΟΜΑΘΗΜΑΤΙΚΩΝ-ΒΙΟΜΕΤΡΙΑΣ

ΕΡΓΑΣΤΗΡΙΟ ΒΙΟΜΑΘΗΜΑΤΙΚΩΝ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΛΑΡΙΣΑ

ΕΙΣΑΓΩΓΗ

Τα ερεθίσματα για την προετοιμασία αυτών των εργαστηριακών σημειώσεων ήταν η εισαγωγή του Μαθήματος Επιλογής Βιομετρία-Βιοπληροφορική στο προπτυχιακό πρόγραμμα σπουδών στο Τμήμα Ιατρικής του Π.Θ. και η εισαγωγή του Μαθήματος Βιοπληροφορική στο Μεταπτυχιακό Πρόγραμμα Σπουδών στο ίδιο Τμήμα. Η εισαγωγή αυτών των μαθημάτων ήταν μετά από πρωτοβουλία των καθηγητών Ν. Σταθάκη, Π. Μολυβδά και Α. Γερμενή.

Πρέπει να τονισθεί ότι στην διεθνή βιβλιογραφία υπάρχουν πολύ λίγα βιβλία στο αντικείμενο και κανένα από αυτά δεν προσφέρει ένα απλό οδηγό στα διάφορα πεδία του αντικειμένου της Βιοπληροφορικής.

Οι σημειώσεις αποτελούνται από δύο ενότητες: Θεωρία και πρακτική εφαρμογή της θεωρίας μέσα από απλά βήματα με οδηγό από ιστοσελίδες. Η συγγραφή των σημειώσεων βασίσθηκε στα βιβλία που δίνονται στο τέλος ως Βιβλιογραφία, και ειδικά στο βιβλίο της πρώην συναδέλφου μου στο National Institute for Medical Research, Medical Research Council, U.K., Cristine Orengo.

Η θεματολογία επιλέχθηκε μετά από συζητήσεις με τον φίλο και συνεργάτη Axel Kowald.

Αρχικά το κείμενο είχε γραφτεί στα Αγγλικά, το κύριος μέρος της μετάφρασης το επιμελήθηκε η Βίκυ Σκαρμεά, επίσης βοήθησε η Μ. Χαϊντις.

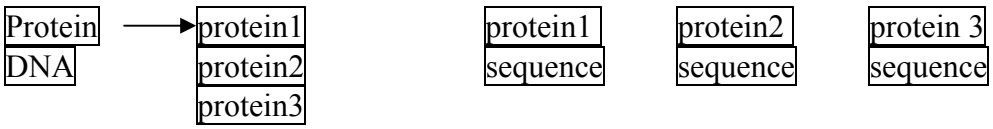
ΠΕΡΙΕΧΟΜΕΝΑ

Σχέση Βιοπληροφορικής και Βιομετρίας

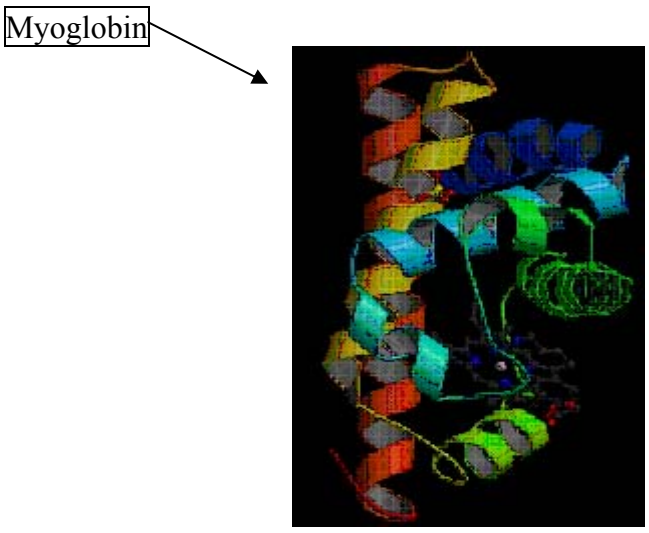
Η βιοπληροφορική είναι η υπολογιστική διαχείριση και στατιστική ανάλυση βιολογικών αλληλουχιών (DNA, RNA, πρωτεΐνες) και δεδομένων που αναφέρονται σε τρισδιάστατες απεικονίσεις πρωτεϊνών.
Η Βιοπληροφορική ασχολείται με:
Τη δημιουργία βάσεων δεδομένων για την αποθήκευση και διαχείριση μεγάλου όγκου δεδομένων.
Την ανάπτυξη αλγορίθμων και βιομετρικών μεθόδων για τον προσδιορισμό σχέσεων μεταξύ αλληλουχιών ή δομών που ανήκουν σε μεγάλα σύνολα δεδομένων.
Την στατιστική ανάλυση και ερμηνεία των διαφορετικών τύπων δεδομένων (DNA, RNA, πρωτεϊνική αλληλουχία και δομή).

Βάσεις Δεδομένων

Μια Βάση Δεδομένων (ΒΠ) είναι μια υπολογιστική αποθήκη από δεδομένα που παρέχει ένα προτυποποιημένο τόπο ανάκτησης, προσθήκης, διαγραφής και αλλαγής δεδομένων.
Υπάρχουν δύο μεγάλες κατηγορίες βάσεων πληροφοριών : οι σχεσιακές και οι αντικειμενοστραφείς.
Οι σχεσιακές ΒΠ τοποθετούν τα δεδομένα σε πίνακες όπου κάθε γραμμή αναφέρεται σε συγκεκριμένο δεδομένο και κάθε στήλη αναφέρεται σε συγκεκριμένο χαρακτηριστικό των δεδομένων. Για τους πίνακες αυτούς δημιουργούνται δείκτες και συσχετίζονται μεταξύ τους, έτσι ώστε κάθε δεδομένο στη ΒΠ να έχει ένα μοναδικό σύνολο χαρακτηριστικών που να το προσδιορίζουν.



Η αντικειμενοστραφής ΒΠ αποτελείται από αντικείμενα (γονίδια ή πρωτεΐνες) που το καθένα έχει ένα συσχετιζόμενο σύνολο από τυποποιημένα εργαλεία για ανάλυση και αναπαράσταση του αντικειμένου και ένα σύνολο από χαρακτηριστικά όπως ένα όνομα που να το προσδιορίζει ή μια παραπομπή.



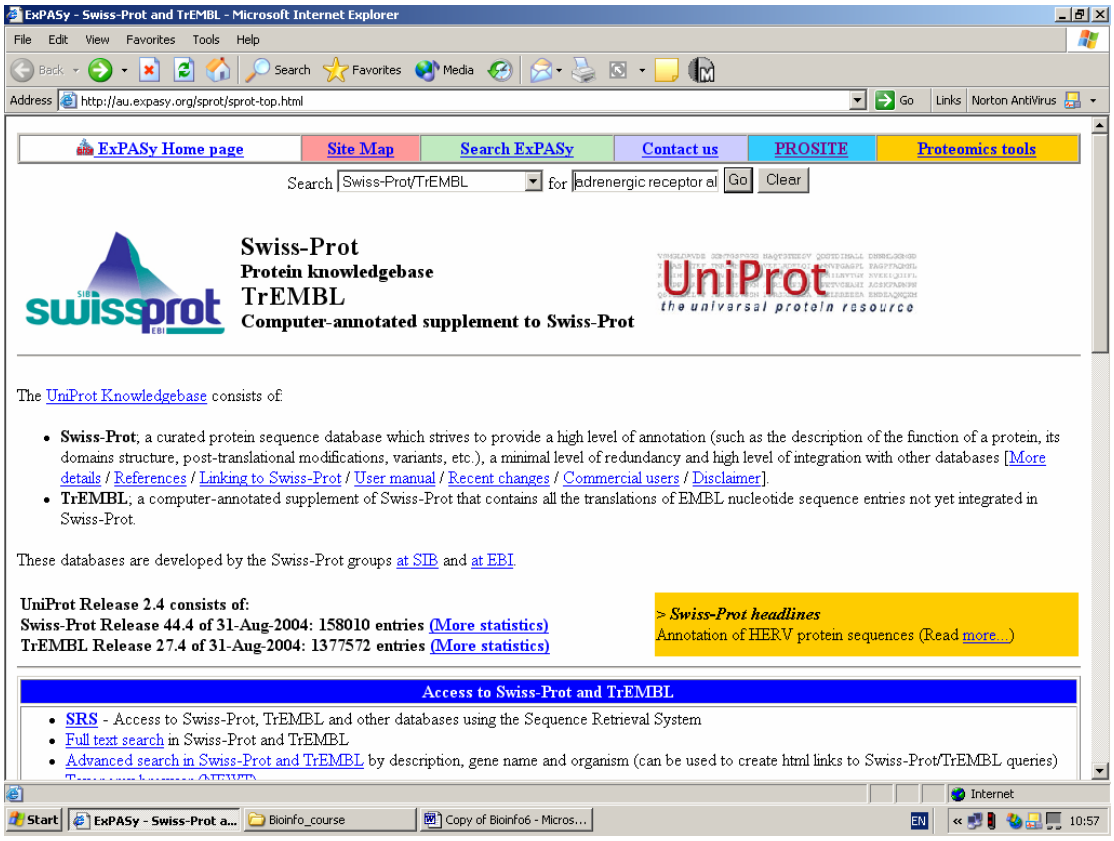
Το Διαδίκτυο
Το Διαδίκτυο είναι ένα παγκόσμιο δίκτυο υπολογιστών που συνδέει δημόσια, εκπαιδευτικά και ιδιωτικά ιδρύματα. Η μετάδοση της πληροφορίας επιτυγχάνεται με τη βοήθεια ενός πρωτοκόλλου γνωστού ως TCP/IP, το οποίο προσδιορίζει τον τρόπο με τον οποίον: 1) τα δεδομένα χωρίζονται σε πακέτα, 2) τα πακέτα φτάνουν στον προσδιορισμό τους και 3) τα πακέτα συναρμολογούνται Το είδος αυτής της επικοινωνίας επιτρέπει διαφορετικά είδη υπολογιστών και λειτουργικών συστημάτων το «μιλήσουν» μεταξύ τους με έναν κοινό τρόπο.

Διευθύνσεις IP
Για να μπορέσουν οι υπολογιστές σε ένα δίκτυο να επικοινωνήσουν, δίνεται σε κάθε υπολογιστή ένας μοναδικός αριθμός (IP διεύθυνση), ο οποίος κωδικοποιείται με μια μορφή δεκαδικής τελείας (π.χ. ένας υπολογιστής στο Διαδίκτυο μπορεί να έχει την IP Διεύθυνση 147.30.32.50). Επειδή η μορφή διευθύνσεων δεν είναι και πολύ φιλική στον άνθρωπο, σε κάθε τέτοια διεύθυνση έχει δοθεί ένα όνομα, το οποίο προσδιορίζει έναν συγκεκριμένο υπολογιστή, τον τόπο που ο υπολογιστής αυτός βρίσκεται, και τέλος το διαδικτυακό μέρος στο οποίο ανήκει αυτός ο τόπος (η παραπάνω διεύθυνση αντιστοιχεί στο όνομα ebi.ac.uk : Το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (European Bioinformatics Institute) είναι ένας ακαδημαϊκός τόπος του ευρύτερου διαδικυακού τόπου UK).

Παγκόσμιος Ιστός (World Wide Web-www)
Ο παγκόσμιος ιστός είναι το πιο προηγμένο σύστημα πληροφόρησης που υπάρχει στο Διαδίκτυο.
Η μετάδοση-μεταβίβαση της πληροφορίας στο www γίνεται με την βοήθεια ενός περιηγητή ιστοσελίδων (browser). Ο πιο γνωστός και πιο διαδεδομένος τέτοιος περιηγητής ιστοσελίδων, είναι ο Internet Explorer. Ο παγκόσμιος ιστός επιτρέπει την επισκόπηση ιστοσελίδων με πληροφορίες πολυμέσων (multimedia) όπως κείμενο, εικόνες κ.τ.λ. που είναι σε μορφή υπερκειμένου (hypertext). Στο υπερκείμενο μερικές λέξεις τονίζονται με διαφορετικό χρώμα και ονομάζονται υπέρ-σύνδεσμοι (hyperlinks). Επιλέγοντας έναν τέτοιο υπέρ-σύνδεσμο, ο περιηγητής ιστοσελίδων προσπελαύνει μια καινούργια ιστοσελίδα με υπερκείμενο (ή μια ΒΠ με μια διεπαφή υπερκειμένου), η οποία μπορεί να βρίσκεται σε κάποιον άλλον υπολογιστή που είναι συνδεδεμένος στο Διαδίκτυο.

Παγκόσμιος Προσδιοριστής Πόρων (ΠΠΠ) (URL)
Ο Παγκόσμιος Ιστός βασίζεται στο ότι κάθε ιστοσελίδα υπερκειμένου έχει έναν Παγκόσμιο Προσδιοριστή Πόρων (Uniform Resource Locator-URL). Ο ΠΠΠ περιέχει διάφορα μέρη, τα οποία προσδιορίζουν το πρωτόκολλο επικοινωνίας (hhtp – hypertext transfer protocol), τον διακομιστή ιστοσελίδων, τον κατάλογο και τέλος την ιστοσελίδα. Για παράδειγμα :
Ο http://www.ebi.ac.uk/hinxton/hinxton.html προσδιορίζει σαν πρωτόκολλο επικοινωνίας το http, σαν Διακομιστή Ιστοσελίδων το (ebi) και σαν κατάλογο το hinxton και τέλος σαν ιστοσελίδα την hinxton.html.

Ένα άλλο παράδειγμα ιστοσελίδας που είναι η διεπαφή μιας ΒΠ για πρωτεΐνες και μπορούμε να προσπελάσουμε με την βοήθεια του Internet Explorer βρίσκεται στο http://au.expasy.org/sprot/sprot-top.html. Επιλέξτε το "at EBI" που είναι ένας υπεσύνδεσμος και επιτρέπει να συνδεθούμε στην ιστοσελίδα του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής.



Διαδικτυακοί Τόποι Βιοπληροφορικής

Τρεις τόποι για να αρχίσει κανείς την αναζήτηση βιοπληροφοριών στο Διαδίκτυο είναι οι εξής :

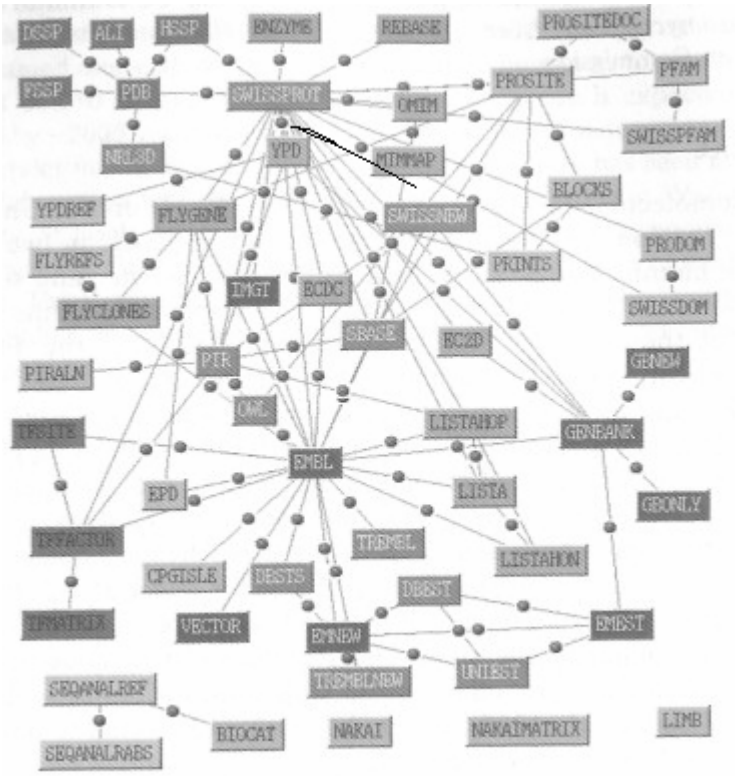
http://www.expasy.ch/: The ExPASy (Expert Protein Analysis System) Molecular Biology Server. Ο συγκεκριμένος διαδικτυακός τόπος διατηρείται από το Ελβετικό Ίδρυμα Βιοπληροφορικής. Περιέχει συνδέσμους, ΒΠ και λογισμικό για την ανάλυση πρωτεϊνικών αλληλουχιών και δομών.

http://www.ebi.ac.uk/: The EMBL European Bioinformatics Institute outstation. Ο τόπος αυτός αποτελεί μία συλλογή ΒΠ με βιολογικό περιεχόμενο και παρεμφερές λογισμικό.

http://www.ncbi.nim.nih.gov/: The National Center for Biotechnology Information. Ο τόπος περιέχει συλλογή δημοσίων ΒΠ , εργαλείων για βιοπληροφορική και διαφόρων εφαρμογών. Παρέχει επίσης συνδέσμους σε πολλούς χρήσιμους διαδικτυακούς τόπους και συνδέσμους για λογισμικό βιοπληροφορικής.

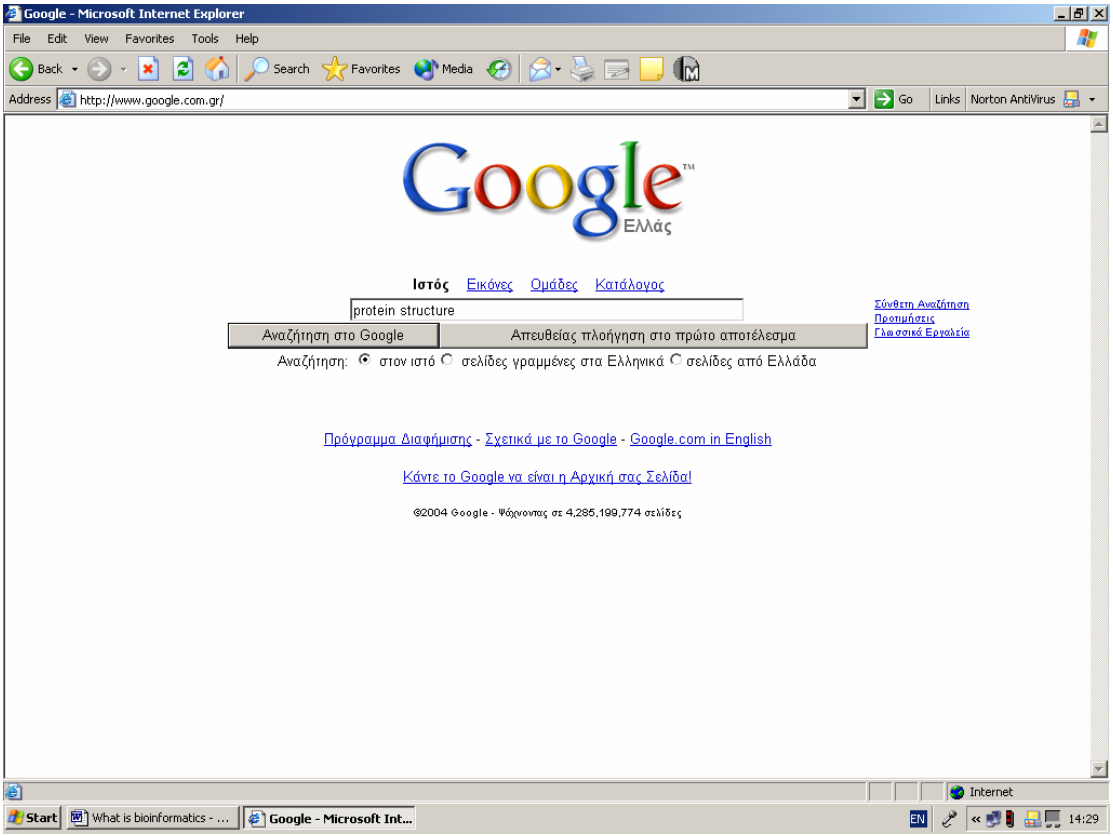
Καθένας από τους παραπάνω διαδικτυακούς τόπους έχει πολλούς συνδέσμους σε άλλους παρόμοιους τόπους βιοπληροφορικής.

Ένα δίκτυο από πηγές βιοπληροφόρισης παρουσιάζεται στο παρακάτω διάγραμμα:



Μηχανές Αναζήτησης (Search Engines)

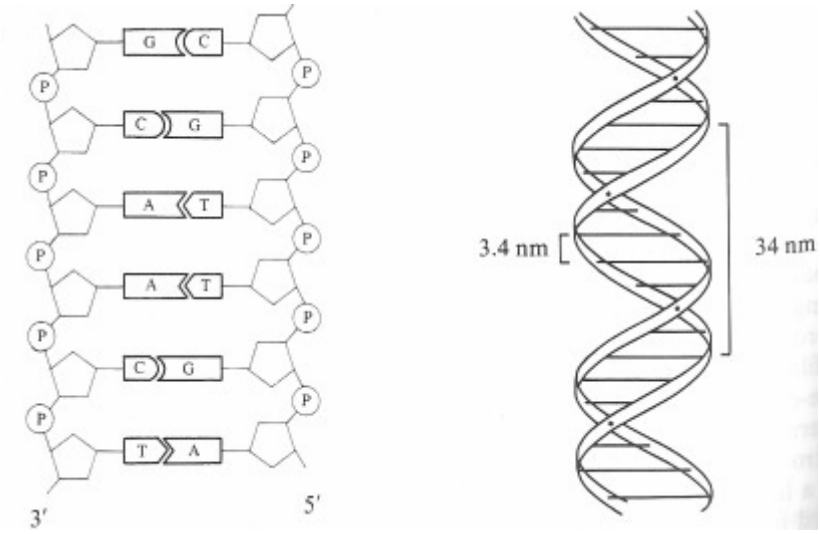
Για να βρούμε οποιοδήποτε είδος πληροφορίας βιολογικού ενδιαφέροντος στο Διαδίκτυο , θα πρέπει να χρησιμοποιήσουμε μηχανές αναζήτησης πληροφορίας. Οι μηχανές αναζήτησης ψάχνουν στις σελίδες του Διαδικτύου για συγκεκριμένες λέξεις κλειδιά και συγκεκριμένες φράσεις. Μια τέτοια μηχανή αναζήτησης είναι το Google. Για να χρησιμοποιήσουμε το Google, γράφουμε www.google.com στο Internet Explorer. Αν πληκτρολογήσουμε πχ τη λέξη κλειδί "protein structure", εμφανίζονται σελίδες που περιέχουν πληροφορίες σχετικά με την λέξη κλειδί.


Google - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.google.com.gr/
Go Links Norton AntiVirus
Google
Ελλάς
Ιστός Εικόνες Ομάδες Κατάλογος
protein structure
Αναζήτηση στο Google
Απευθείας πλοήγηση στο πρώτο αποτέλεσμα
Σύνθετη Αναζήτηση
Προτιμήσεις
Γλωσσικά Εργαλεία
Αναζήτηση: στον ιστό σελίδες γραμμένες στα Ελληνικά σελίδες από Ελλάδα
Πρόγραμμα Διαφήμισης - Σχετικά με το Google - Google.com in English
Κάντε το Google να είναι η Αρχική σας Σελίδα!
©2004 Google - Ψάχνοντας σε 4,285,199,774 σελίδες
Internet
Start
What is bioinformatics - ...
Google - Microsoft Int...
EN
14:29

Βασική Βιολογία

DNA

Το DNA είναι ένα μεγάλο, γραμμικό μόριο του οποίου η δομή αποτελείται από δύο αλυσίδες τυλιγμένες μεταξύ τους σαν δύο κλώνους ενός νήματος. Κάθε αλυσίδα αποτελείται από μια βάση σακχάρων, στην οποία είναι προσκολλημένα μια αλληλουχία νουκλεοτιδίων ή βάσεων. Υπάρχουν τέσσερις βάσεις: Αδενίνη (A), Κυτοσίνη (C), Γουανίνη (G) και Θειαμίνη (T).
Η αλληλουχία των βάσεων στις δύο αλυσίδες είναι συμπληρωματική (A με T και C με G). Αυτό το συμπληρωματικό ζευγάρωμα των βάσεων είναι υπεύθυνο για τον χημικό δεσμό που κρατά τις αλυσίδες μαζί σε ένα διπλό έλικα και είναι βασικό στην αντιγραφή του DNA κατά την διαίρεση του κυττάρου.



Συσχετισμένη με κάθε αλυσίδα, είναι και η κατεύθυνση, στην οποία τα μόρια που είναι υπεύθυνα να διαβάζουν την πληροφορία της έλικας πρέπει να κινηθούν. Η κατεύθυνση αυτή ονομάζεται 5' – 3' κατεύθυνση της έλικας.

RNA

Το RNA διαφέρει από το DNA στο ότι είναι μόριο με μία μόνο αλυσίδα και στο γεγονός ότι η Θειαμίνη (T) αντικαθίσταται από την Ουρασίλη (U).

Γενετικός Κώδικας

Ο γενετικός κώδικας προσδιορίζει τον τρόπο με τον οποίον η αναγκαία πληροφορία για να περιγραφεί μια πρωτεΐνη είναι κωδικοποιημένη στο DNA. Μπορούμε να φανταστούμε το DNA σαν ένα κομμάτι από κείμενο, γραμμένο με ένα αλφάβητο τεσσάρων λέξεων : A,C,G and T.Οι λέξεις σε αυτό το κείμενο είναι διαδοχικές μη-επικαλυπτόμενες τριπλέτες βάσεων, που ονομάζονται κωδικόνια (codons). Κάθε κωδικόνιο κρυπτογραφεί ένα αμινοξύ (amino acid) ή πεπτίδιο (peptide), και η διαδοχή των κοδωνίων στο DNA προσδιορίζει μια συγκεκριμένη πολυπεπτιδική αλυσίδα η οποία αναφέρεται σε μια συγκεκριμένη πρωτεΐνη ή μέρος μιας πρωτεΐνης. Η μεταφρασμένη αλληλουχία DNA είναι μια πρωτεΐνη.
Υπάρχουν 20 αμινοξέα στην φύση και 64 διαφορετικά κωδικόνια. Αυτό σημαίνει ότι πολλά από τα αμινοξέα αντιπροσωπεύονται από παραπάνω του ενός κωδικονίου. Ο γενετικός κώδικας φαίνεται παρακάτω :

	T		C		A		G		
T	TTT	Phe	TCT	Ser	TAT	Try	TGT	Cys	T
	TTC		TCC		TAC		TGC		C
	TTA	Leu	TCA		TAA	**Stop**	TGA	**Stop**	A
	TTG		TCG		TAG		TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lys	AGA	Arg	A
	ATG	**Met**	ACG		AAG		AGG		G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glu	GGA		A
	GTG		GCG		GAG		GGG		G

Εάν πάρουμε υπ' όψιν μας την αλληλουχία

5' . . . ACGTTGGGACTAAGTC . . . 3'

η οποία θα μπορούσε να διαιρεθεί σε κωδικόνια και να μεταφράσθεί ως ακολούθως:

5' . . . – ACG – TTG – GGA – CTA – AGT - C . . . 3'
. . . Thr Leu Gly Leeu Ser . . .

Όμως υπάρχουν και άλλα πιθανά σημεία για την έναρξη της μετάφρασης, διαμορφώνοντας έτσι διαφορετικούς τρόπους ανάγνωσης της αλληλουχίας:

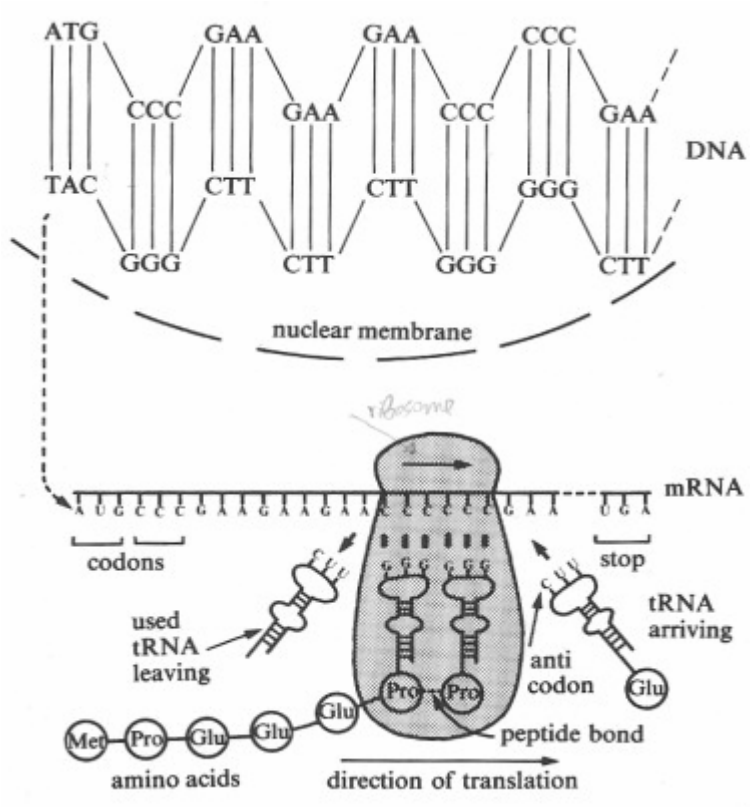
5' . . . AC – GTT – GGG – ACT – AAG - TC . . . 3'
. . . Val Gly Thr Lys . . .

Υπάρχει λοιπόν η ανάγκη να οριστεί η αρχική θέση ανάγνωσης για μια αλληλουχία. Και αυτό γίνεται μαρκάροντας την αντιγραφή με το ειδικό κωδικόνιο ATG. Και κατά επέκταση, υπάρχουν και ειδικά κωδικόνιο τερματισμού, TAA, TAG and TGA.

Αποτελέσματα αλλαγών στην κωδικοποίηση του DNA
Αλλαγές στην κωδικοποίηση που δεν αλλάζουν την θέση ανάγνωσης, δημιουργούν πάρα πολύ μικρές αλλαγές στην μεταφρασμένη αλληλουχία της πρωτεΐνης. Για παράδειγμα, αντικαθιστώντας μια μόνη βάση από μια άλλη διαφορετική βάση, μπορεί, στην χειρότερη περίπτωση, να αλλάξει μόνο το αμινοξύ που κωδικοποιείται από το συγκεκριμένο κωδικόνιο όπου η αλλαγή έγινε. Όμως η απομάκρυνση ή η προσθήκη μιας μόνο βάσης μπορεί να αλλάξει την αρχική θέση ανάγνωσης.

Σύνθεση πρωτεΐνης

Σύνθεση πρωτεΐνης, είναι η διαδικασία συνδυασμού αμινοξέων για την δημιουργία πολυπεπτιδίων. Η διαδικασία μετάφρασης μιας αλληλουχίας DNA με την βοήθεια των mRNA σε μια πολυπεπτιδική αλυσίδα μέσα σε ένα ριβόσωμα με την δέσμευση αμινοξέων που μεταφέρονται από τα tRNA φαίνεται παρακάτω:



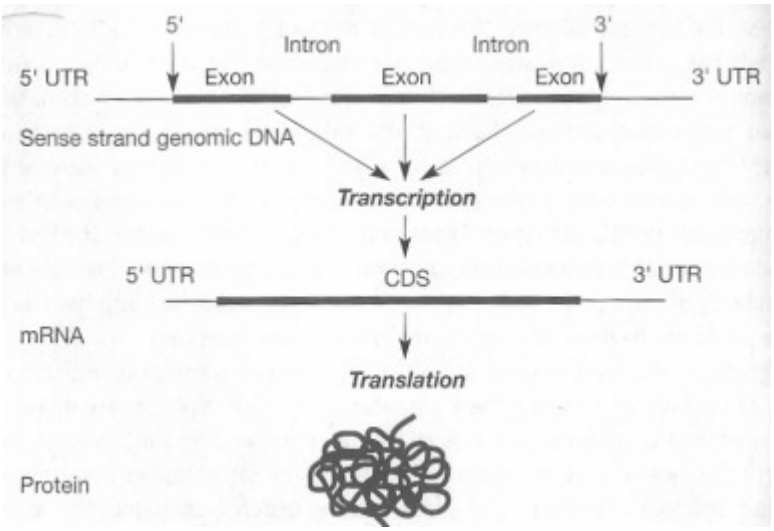
Όταν έχει γίνει η σύνθεση του πολυπεπτιδίου, το εν λόγω πολυπεπτίδιο μπορεί να υποστεί μετά-μεταφρασμένες μετατροπές (post-translational modification) , όπως να αναδιπλωθεί ή να συσχετιστεί με άλλα πολυπεπτίδια για να σχηματίσει μια λειτουργική πρωτεΐνη.

Γονίδια

Μια αλληλουχία DNA που κωδικοποιεί κάποια πρωτεΐνη, ονομάζεται γονίδιο. Ένα γονίδιο μπορεί να βρίσκεται σε οποιονδήποτε από τους δύο κλώνους της αλυσίδας του DNA.

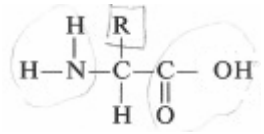
Το σύνολο του DNA ονομάζεται γονιδίωμα, και στους ευκαριωτικούς οργανισμούς το περισσότερο μέρος του DNA (99%) δεν κωδικοποιεί.

Στα γονίδια στους ευκαριώτες μεταξύ των περιοχών που κωδικοποιούνται (τα exons) παρεμβάλλονται περιοχές που δεν κωδικοποιούνται (τα introns).



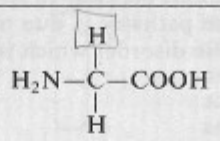
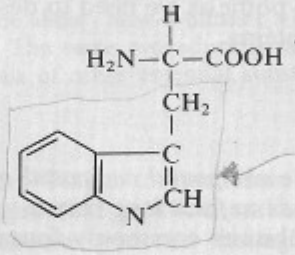
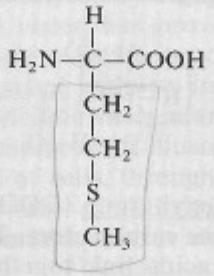
Πρωτεϊνική δομή

Η πρωτεΐνη είναι μια οργανική ένωση που αποτελείται από αμινοξέα, τα οποία συνδέονται μεταξύ τους με πεπτιδικούς δεσμού, για να σχηματίσουν μεγάλες αλυσίδες, που ονομάζονται πολυπεπτίδια. Υπάρχουν 20 διαφορετικά αμινοξέα που συναντούνται στις πρωτεΐνες. Τα αμινοξέα μπορούν να παρασταθούν ως ακολούθως:



Όλα τα αμινοξέα έχουν μία καρβοξυλική ομάδα (COOH) και μια αμινοομάδα (NH2), οι οποίες συνδέονται στην ίδια ανθρακική αλυσίδα. Διαφέρουν μόνο στο μέρος της πλευρικής αλυσίδας (side-chain) που αναπαριστάται με R.

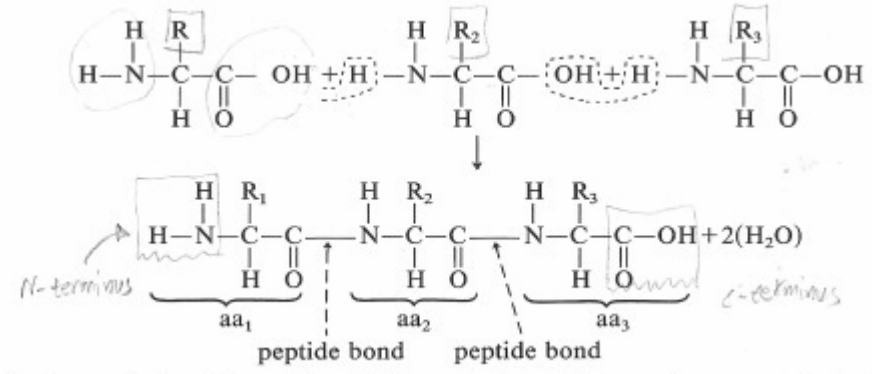
Τα 20 αμινοξέα που βρίσκονται στις πρωτεΐνες είναι τα παρακάτω:

Amino acid	*Abbrevi-ation*	*Examples of structure*
Alanine	Ala	 **Glycine** is the simplest amino acid.
Arginine	Arg	
Asparagine	Asn	
Aspartic acid	Asp	
Cysteine	Cys	
Glutamine	Gln	
Glutamic acid	Glu	 **Tryptophan** has an aromatic side chain.
Glycine	Gly	
Histidine	His	
Isoleucine	Ile	
Leucine	Leu	
Lysine	Lys	
Methionine	Met	 **Methionine** has a sulphur-containing side chain.
Phenylalanine	Phe	
Proline	Pro	
Serine	Ser	
Threonine	Thr	
Tryptophan	Trp	
Tyrosine	Tyr	
Valine	Val	

Υπάρχει επίσης μια και μια πρότυπη κωδικοποίηση με ένα γράμμα για τα αμινοξέα:

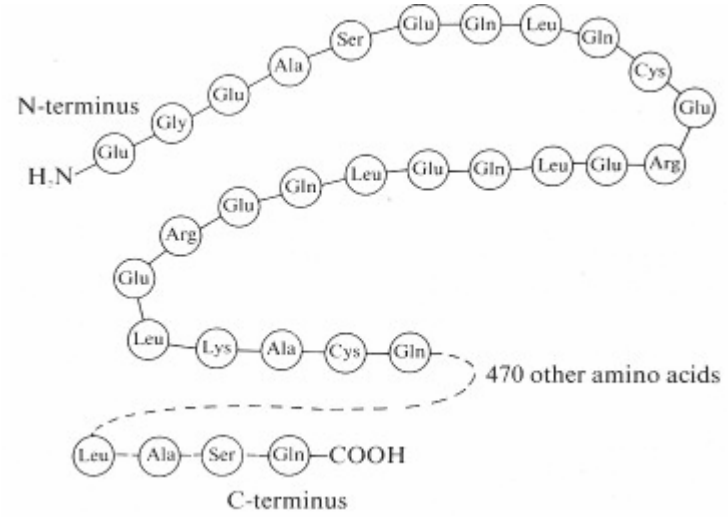
A=Ala	G=Gly	M=Met	S=Ser
C=Cys	H=his	N=Asn	T=Thr
D=Asp	I=Ile	P=Pro	V=Val
E=Glu	K=Lys	Q=Gln	W=Trp
F=Phe	L=Leu	R=Arg	Y=Tyr

Τα αμινοξέα ενώνονται μεταξύ τους με μία αντίδραση ανάμεσα στη καρβοξυλομάδα και στη αμινομάδα για να δημιουργήσουν έναν πεπτιδικό δεσμό. Μόρια νερού δημιουργούνται κατά την αντίδραση. Η διαδικασία αυτή δημιουργεί μεγάλες αλυσίδες πεπτιδίων (πολυπεπτίδια).



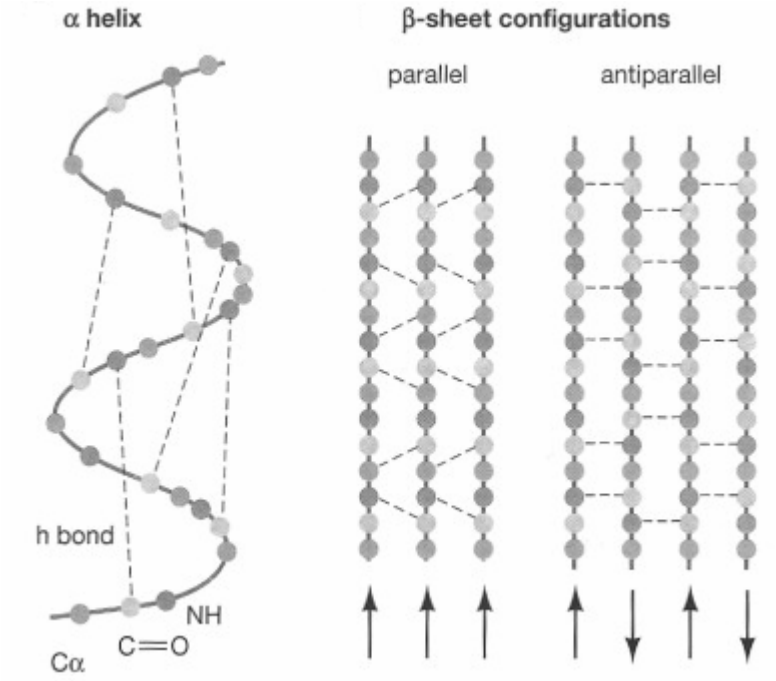
Κάθε πολυπεπτίδιο έχει μια ελεύθερη αμινομάδα στο ένα μέρος (N-terminus) και μια καρβοξυλομάδα στο άλλο μέρος (C-terminus).
Οι πρωτεΐνες διαφέρουν στη σύνθεση και στη ακολουθία των αμινοξέων τους.

Κάθε πρωτεΐνη έχει μια ιδιαίτερη αλληλουχία, η οποία είναι ουσιώδους σημασίας για τη δομή και λειτουργία της. Ένα παράδειγμα μιας ακολουθίας είναι η:



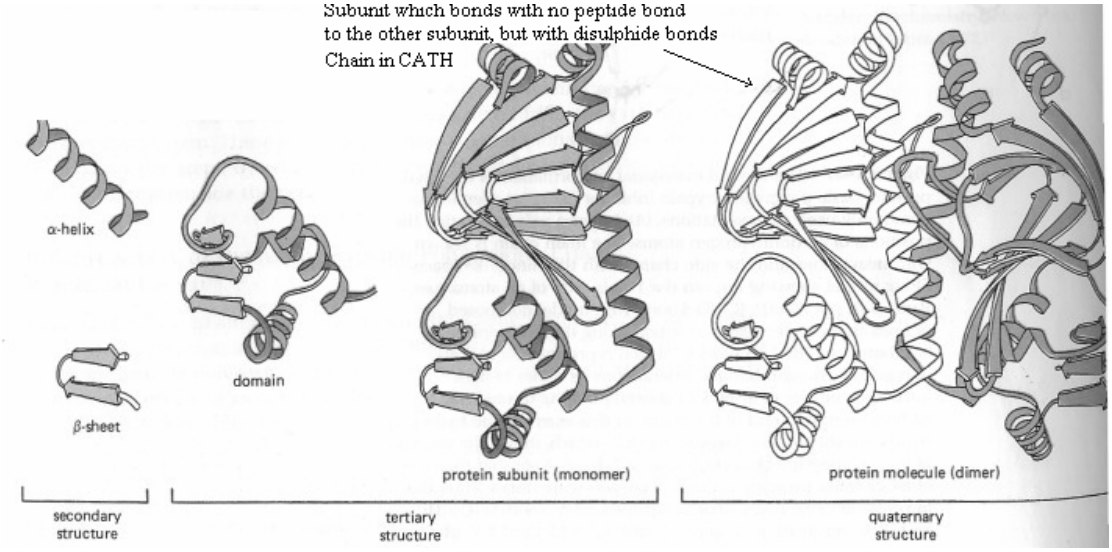
Ο προσδιορισμός της αλληλουχίας ενός αμινοξέους ονομάζεται αλληλούχιση (sequencing).

Επίπεδα της πρωτεϊνικής δομής
Πρωτοταγής δομή είναι η γραμμική αλληλουχία αμινοξέων σε ένα μόριο πρωτεΐνης.



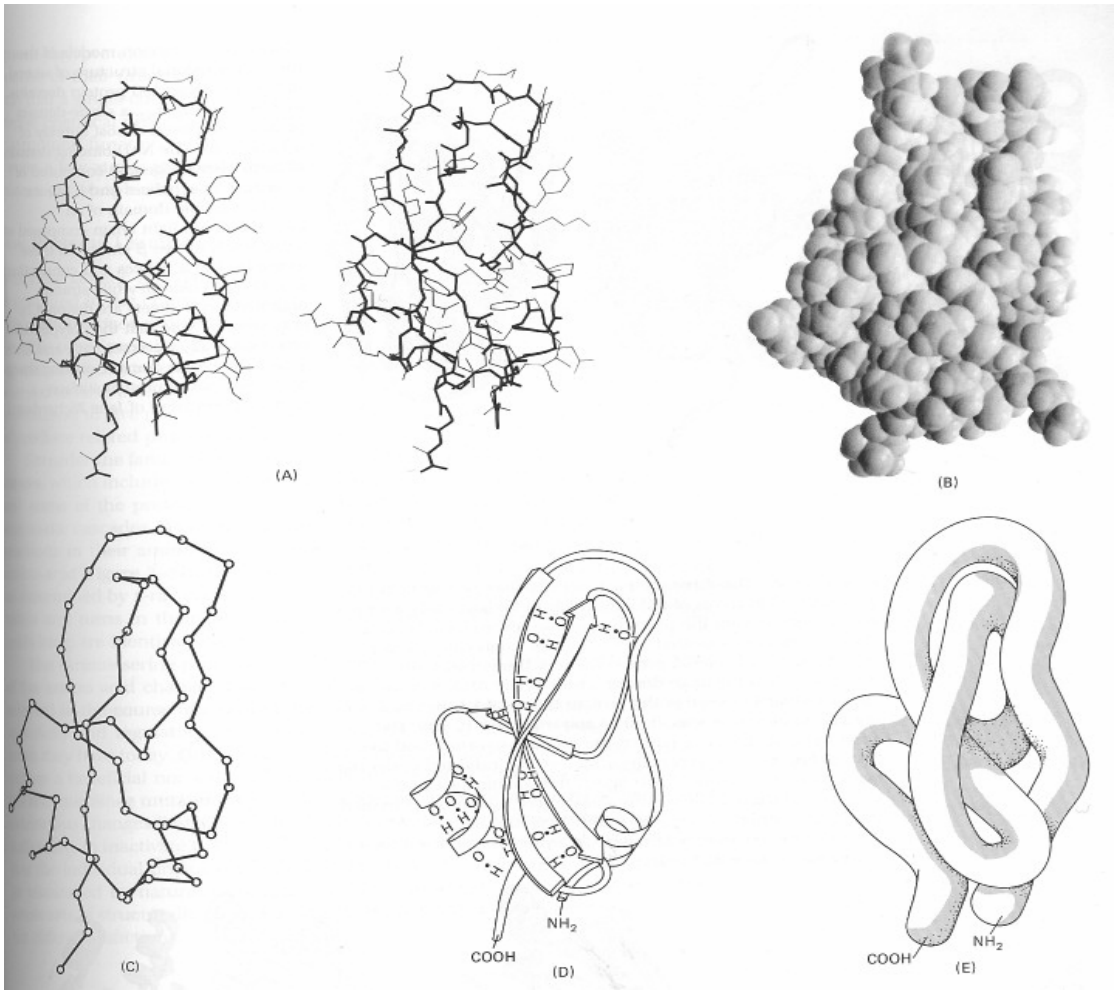
Αμινοξέα, τα οποία είναι δίπλα το ένα στο άλλο στην αλυσίδα, αλληλεπιδρούν (αλληλεπιδράσεις δεσμών υδρογόνου) με πολλαπλούς τρόπους: αυτό έχει σαν

συνέπεια το ένα μέρος του γραμμικού πολυπεπτιδίου να γίνει ελικοειδές, σχηματίζοντας μια a-έλικα (a-helix) και το άλλο μέρος να σχηματίσει ένα πλέγμα με γραμμές που έχουν μία κατεύθυνση (b-sheet). Αυτές οι αλληλεπιδράσεις δημιουργούν την δευτεροταγή δομή.
Τριτοταγής δομή είναι η ολική δίπλωση μιας πρωτεϊνικής αλληλουχίας, η οποία σχηματίζεται από το πακετάρισμα των δευτεροταγών (δισουλφιδικοί θειούχοι δεσμοί ανάμεσα σε αμινοξέα που περιέχουν θείον σταθεροποιούν το μόριο στην τρισδιάστατη δομή του).
Τεταρτοβάθμια δομή είναι η διάταξη διαφορετικών πολυπεπτιδικών αλυσίδων (υποομάδες-subunits) σε ένα μόριο πρωτεΐνης.



Τα επίπεδα της πρωτεϊνικής δομής προσδιορίζονται από την θέση των αμινοξέων στην κύρια αλυσίδα, η σωστή αλληλουχία είναι ζωτικής σημασίας εάν η πρωτεΐνη πρέπει να λειτουργήσει με τον σωστό τρόπο.
Ο τομέας (domain) μιας πρωτεΐνης είναι ο συνδυασμός από a-helices και b-sheets, τα οποία ενώνονται μεταξύ τους για να σχηματίσουν συμπαγές διπλωμένες σφαιρικές μονάδες.
Οι υποομάδες είναι πολυπεπτίδια που ενώνονται με άλλες υποομάδες με ασθενείς δεσμούς (δισουλφιδικούς δεσμούς).

Τρισδιάστατη απεικόνιση πρωτεϊνών
Η τρισδιάστατη απεικόνιση πρωτεϊνών μπορεί να απεικονιστεί με διαφόρους τρόπους. Για παράδειγμα, η πρωτεΐνη "basic pancreatic trypsin inhibitor (BPTI)" μπορεί να αναπαρασταθεί με :
Γραμμές, όπου η κύρια αλυσίδα αναπαριστάται με έντονες γραμμές και η πλευρικές αλυσίδες με λεπτές γραμμές
Μοντέλο συμπλήρωσης κενών (Space-filling model)
Μοντέλο αξονικών συμπλεγμάτων (backbone wire model), όπου κάθε άτομο άνθρακα συνδέεται με γραμμές κατά μήκος της πολυπεπτιδικής αλυσίδας.
Μοντέλο Κορδέλας (Ribbon model), το οποίο αναπαριστά όλες τις περιοχές των υδρογονικών δεσμών ή σαν έλικες (α-έλικες) ή γραμμές με κατεύθυνση (b-sheets) που δείχνουν προς το καβοξυλιακό μέρος του αλυσίδας.
Μοντέλο του Λουκάνικου (Sausage model)



Συγκεκριμένοι συνδυασμοί των a-helices και b-sheets που δημιουργού μία σφαιρική δομή και υπάρχουν επαναλαμβανόμενες σε πολλές πρωτεΐνες που δεν σχετίζονται μεταξύ τους ονομάζονται μοτίβα (motifs). Για παράδειγμα, το beta-alpha-beta μοτίβο, το οποίο συναντάται σε πολλές διαφορετικές πρωτεΐνες.

Πληροφοριακά συστήματα Γονιδιώματος

GenBank

Η GenBank είναι μία βάση δεδομένων με πληροφορίες DNA. Την βάση την διαχειρίζεται το National Center for Biotechnology Information (NCBI).). Η GenBank χωρίζεται σε 17 διακριτές κατηγορίες (divisions) ανάλογα με το είδος της αλληλουχίας DNA που περιέχουν:

Division	sequence subset	Division	sequence subset
PRI	primate	PHG	Bacteriophage
ROD	rodent	SYN	synthetic
MAM	other mammalian	UNA	unannotated
VRT	other vertebrate	EST	expressed sequence tags
INV	invertebrate	PAT	patent
PLN	plant, fungal, algal	STS	sequence tagged sites
BCT	bacterial	GSS	genome survey sequences
RNA	structural RNA	HTG	high throughput genomic sequences
VRL	viral		

Πληροφορίες μπορεί κανείς να ανακτήσει από τη GenBank, χρησιμοποιώντας το ολοκληρωμένο σύστημα ανάκτησης πληροφοριών Entrez.

Δομή των καταχωρήσεων της GenBank

Μια καταχώρηση στη GenBank περιέχει το αρχείο αλληλουχίας, το οποίο περιέχει εκτός από την αλληλουχία καθ'αυτή και διάφορες περιγραφικές πληροφορίες που σχετίζονται με αυτήν.

Κάθε καταχώρηση (entry) επίσης περιλαμβάνει και διάφορες λέξεις-κλειδιά, σχετιζόμενα υπό-κλειδιά και ένα προαιρετικό αρχείο αξιοσημείωτων χαρακτηριστικών(feature table).

Μια GenBank γραμμογράφηση μιας καταχώρηση για την ανθρώπινη cyclooxygenase είναι η ακόλουθη:

```
LOCUS       HUMCYCLOX               3387 bp    mRNA    linear   PRI 31-DEC-1994
DEFINITION  Homo sapiens cyclooxygenase-2 (Cox-2) mRNA, complete cds.
ACCESSION   M90100
VERSION     M90100.1  GI:181253
KEYWORDS    cyclooxygenase-2; prostaglandin synthase.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 3387)
  AUTHORS   Hla,T. and Neilson,K.
  TITLE     Human cyclooxygenase-2 cDNA
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 89 (16), 7384-7388 (1992)
  MEDLINE   92366465
   PUBMED   1380156
COMMENT     Original source text: Homo sapiens umbilical vein cDNA to mRNA.
FEATURES             Location/Qualifiers
     source          1..3387
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /cell_type="endothelial"
                     /tissue_type="umbilical vein"
     gene            1..3387
                     /gene="Cox-2"
     5'UTR           1..97
                     /gene="Cox-2"
     CDS             98..1912
                     /gene="Cox-2"
                     /EC_number="1.14.99.1"
                     /codon_start=1
                     /product="cyclooxygenase-2"
                     /protein_id="AAA58433.1"
                     /db_xref="GI:181254"
                     /translation="MLARALLLCAVLALSHTANPCCSHPCQNRGVCMSVGFDQYKCDC
                     TRTGFYGENCSTPEFLTRIKLFLKPTPNTVHYILTHFKGFWNVVNNIPFLRNAIMSYV
                           …
                     VEVGAPFSLKGLMGNVICSPAYWKPSTFGGEVGFQIINTASIQSLICNNVKGCPFTSF
                     SVPDPELIKTVTINASSSRSGLDDINPTVLLKERSTEL"
     sig_peptide     98..148
                     /gene="Cox-2"
     mat_peptide     149..1909
                     /gene="Cox-2"
                     /product="cyclooxygenase-2"
                     /EC_number="1.14.99.1"
     3'UTR           1913..3387
                     /gene="Cox-2"
     polyA_signal    3369..3374
                     /gene="Cox-2"
ORIGIN
        1 gtccaggaac tcctcagcag cgcctccttc agctccacag ccagacgccc tcagacagca
       61 aagcctaccc ccgcgccgcg ccctgcccgc cgctgcgatg ctcgcccgcg ccctgctgct
      121 gtgcgcggtc ctggcgctca gccatacagc aaatccttgc tgttcccacc catgtcaaaa
      181 ccgaggtgta tgtatgagtg tgggatttga ccagtataag tgcgattgta cccggacagg
      241 attctatgga gaaaactgct caacaccgga atttttgaca agaataaaat tatttctgaa
      301 acccactcca aacacagtgc actacatact tacccacttc aagggatttt ggaacgttgt
      361 gaataacatt cccttccttc gaaatgcaat tatgagttat gtgttgacat ccagatcaca
            …
     3301 tacctgaact tttgcaagtt ttcaggtaaa cctcagctca ggactgctat ttagctcctc
     3361 ttaagaagat taaaaaaaaa aaaaaag
//
```

Η Γραμμή LOCUS είναι το χαρακτηριστικό που υποδουλώνει αλληλουχία λειτουργικότητας (HUMCYCLOX υποδουλώνει ανθρώπινη cyclooxygenase), τον αριθμό βάσεων, την προέλευση της πληροφορίας αλληλουχίας (mRNA), το τμήμα της ΒΠ (PRI) και την ημερομηνία υποβολής των δεδομένων στη ΒΠ.

Η γραμμή DEFINITION περιέχει την περιγραφή της αλληλουχίας.

Η γραμμή ACCESSION είναι ένας μοναδικός κωδικός ο οποίος δίνεται σε κάθε μια καταχώρηση.

Η γραμμή NID περιέχει ένα προσδιορισμό νουκλεοτιδίου, το οποίο επιτρέπει την αλληλουχία να επανεξεταστεί και να συσχετισθεί με το γραμμή LOCUS και τη γραμμή ACCESSION.

Η γραμμή KEYWORDS περιέχει φράσεις που περιγράφουν τα γονίδια και άλλες σχετικές πληροφορίες.

Η γραμμή SOURCE περιέχει πληροφορίες για τον ιστό από τον οποίον τα δεδομένα εξήχθησαν.

Η γραμμή ORGANISM περιγράφει την βιολογική ταξινόμηση του οργανισμού προέλευσης.

Η γραμμή REFERENCE παρέχει παραπομπές στη βιβλιογραφία για την συγκεκριμένη αλληλουχία.

Η γραμμή FEATURES παρέχει ακριβείς πληροφορίες για τα χαρακτηριστικά (Feature Table). Συντεταγμένες παρέχονται για το 5’ μη-μεταφρασμένο μέρος (1-97), για την αλληλουχία κωδικοποίησης (98-1912), για το 3’΄μη-μεταγραφσμένο μέρος (1913-3387), για τη polyadenylation αλληλουχία (3369-3374) κ.τ.λ. Επίσης παρέχονται πληροφορίες για τη μετάφραση της πρωτεΐνης και τις θέσεις διαφόρων πεπτιδίων.

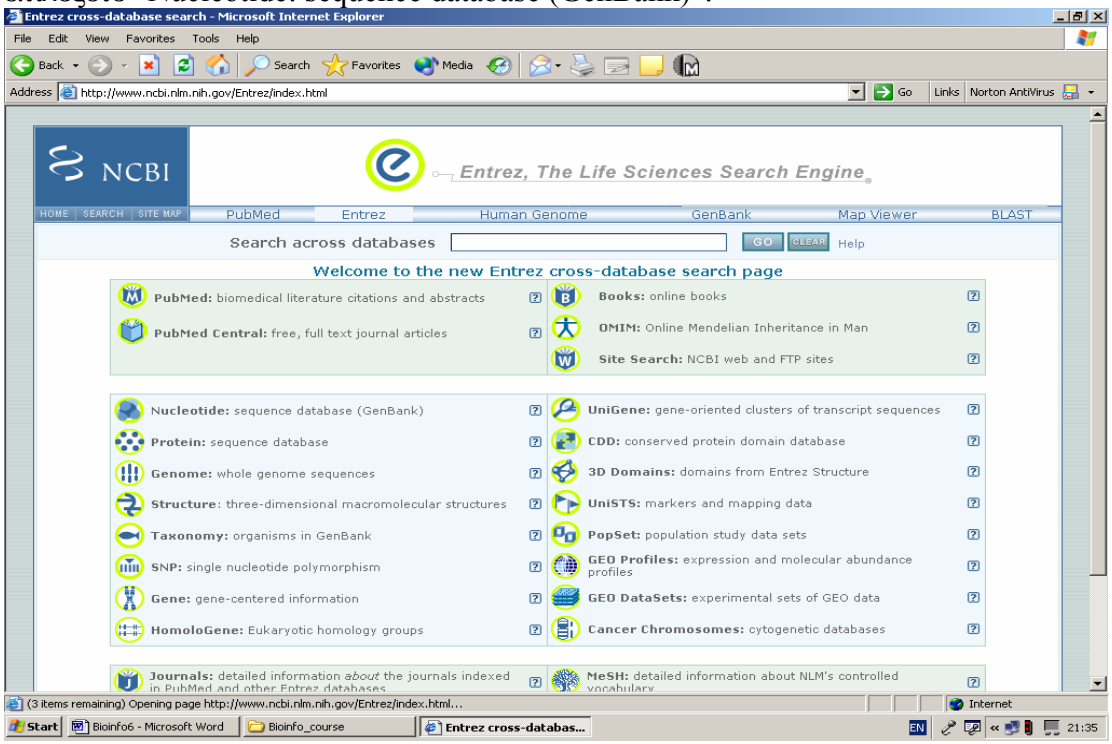
Η γραμμή BASE COUNT δίνει πληροφορίες για την συχνότητα εμφάνισης των διαφορετικών τύπων βάσεων στην αλληλουχία (i.e. 1010 A, 712 C, 633 G, 1032 T).

Η γραμμή ORIGIN σημειώνει την πρώτη βάση της αλληλουχίας στο γονιδίωμα.

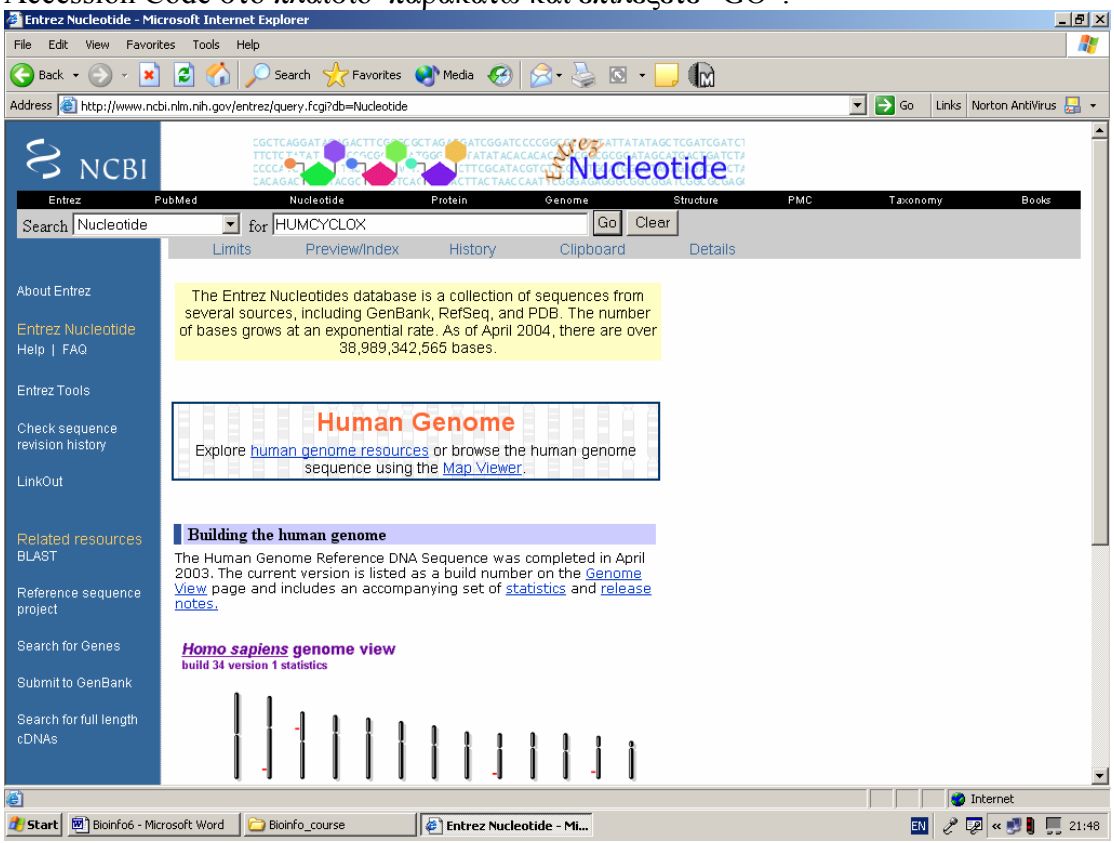
ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

GenBank

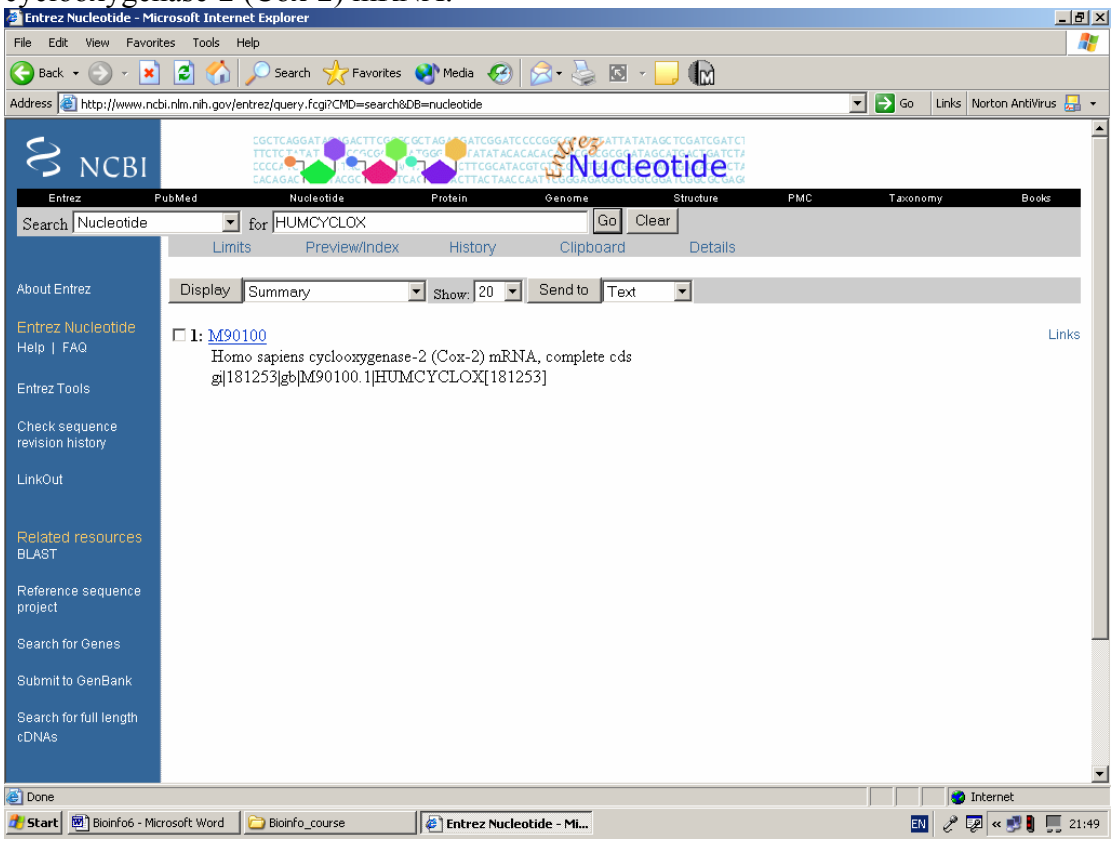
Για να προσπελάσετε τη ΒΠ GenBank, στον Internet Explorer πληκτρολογήστε την ακόλουθη διεύθυνση : www.ncbi.nlm.nih.gov/Entrez/index.html. Στη συνέχεια επιλέξετε "Nucleotide: sequence database (GenBank)".



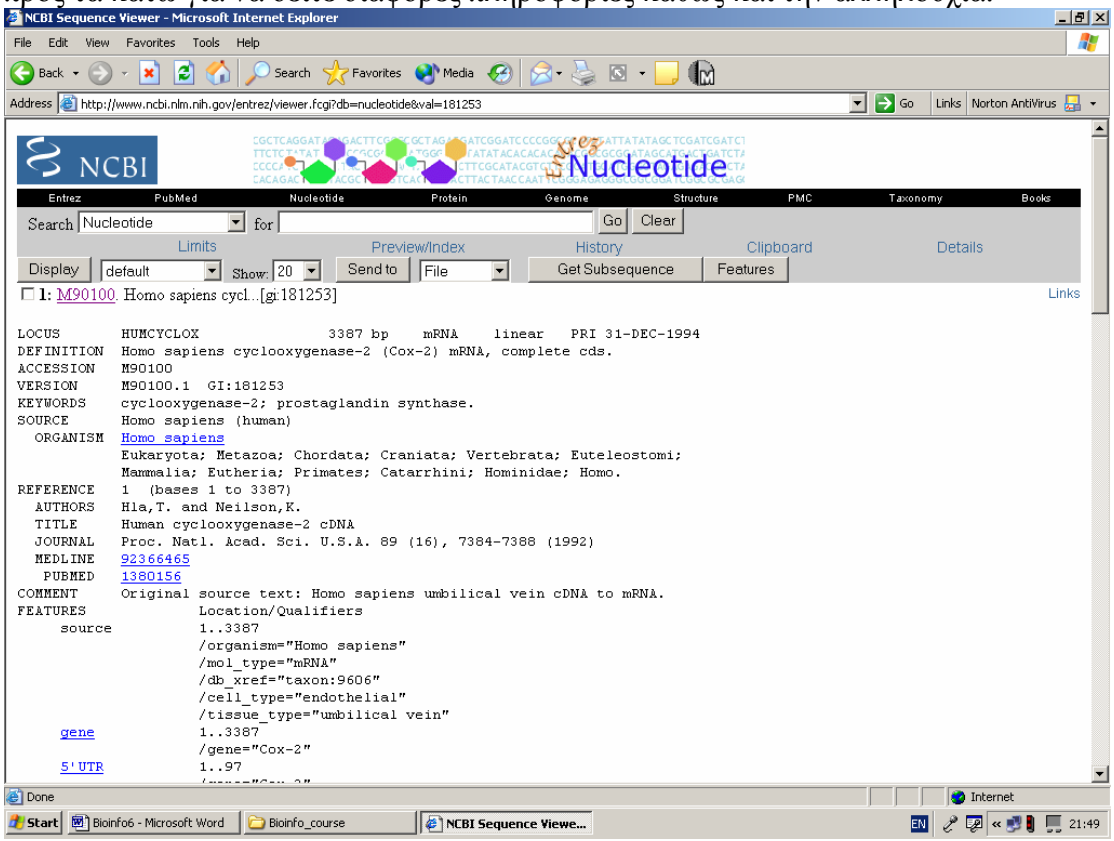
Για να ανακτήσετε πληροφορίες για το HUMCYCLOX, πληκτρολογήστε τον Accession Code στο πλαίσιο παρακάτω και επιλέξετε "GO".

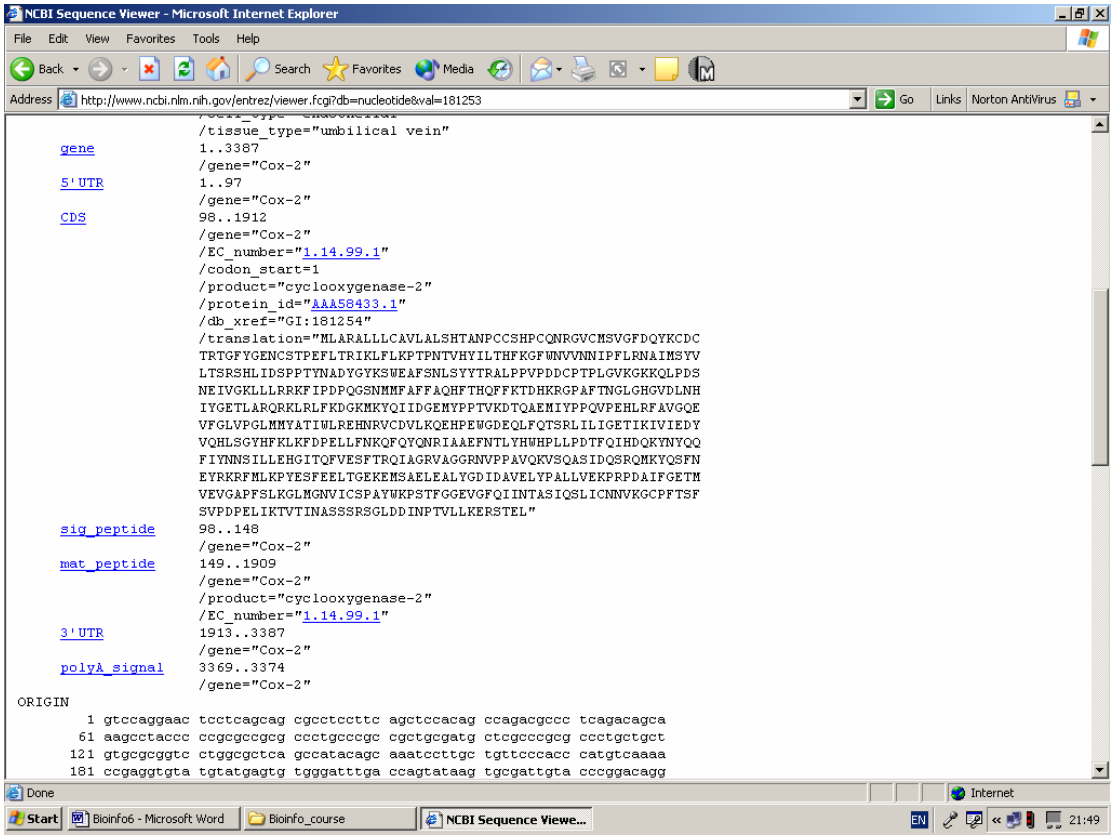


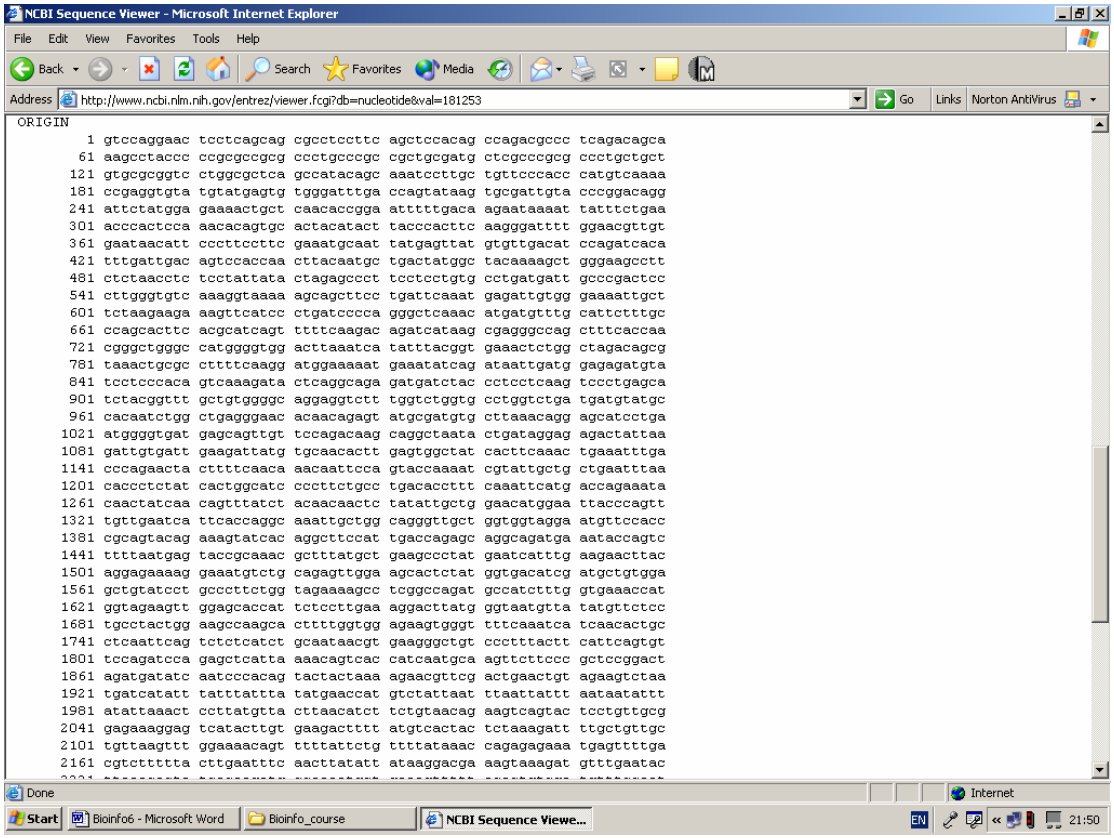
Στη συνέχεια επιλέξετε “M90100”, τον accession code για Homo sapiens cyclooxygenase-2 (Cox-2) mRNA.

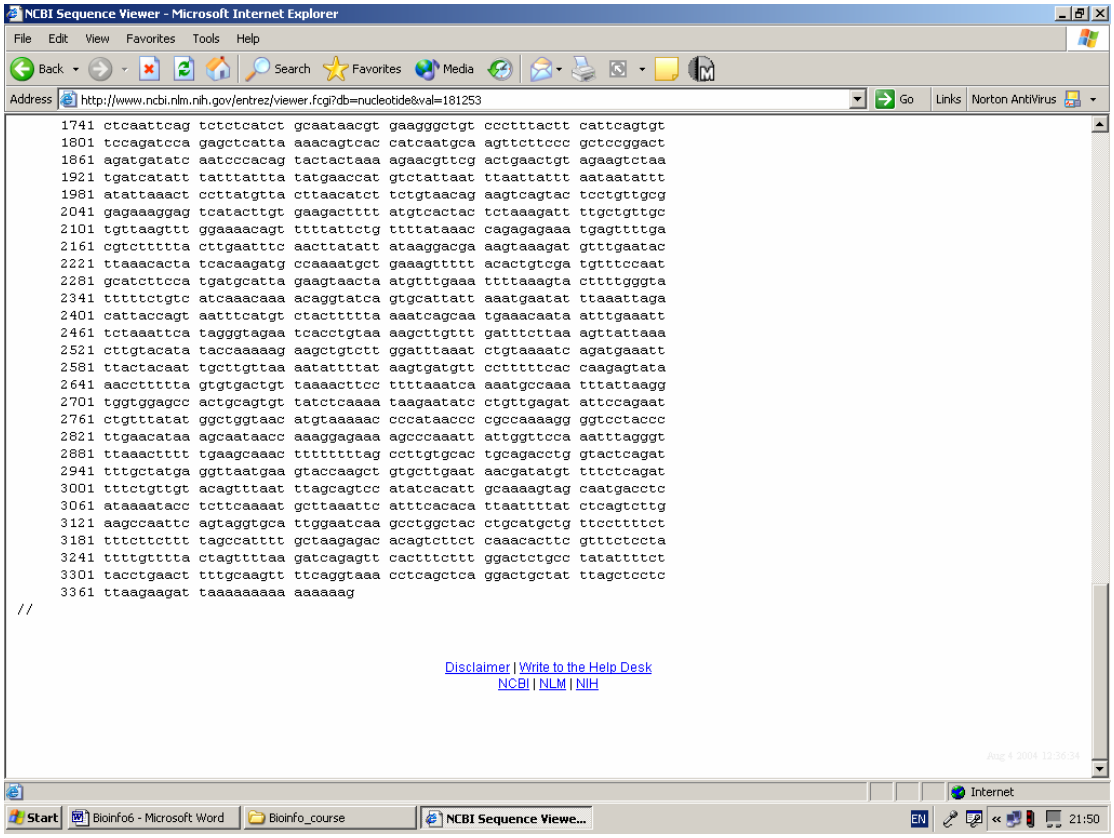


Το σύστημα θα παρουσιάσει την καταχώρηση. Μπορείτε να κυλήσετε την οθόνη προς τα κάτω για να δείτε διάφορες πληροφορίες καθώς και την αλληλουχία.

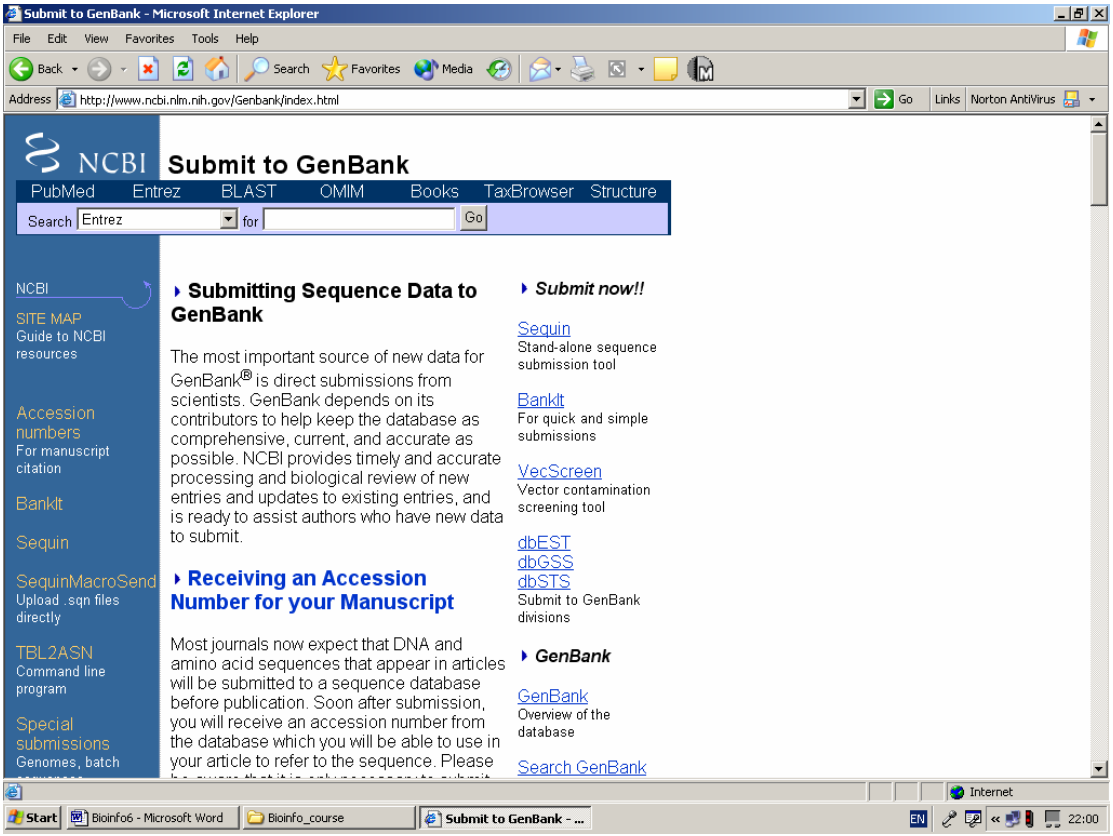



NCBI Sequence Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=181253
Go Links Norton AntiVirus
/tissue_type="umbilical vein"
gene 1..3387
/gene="Cox-2"
5'UTR 1..97
/gene="Cox-2"
CDS 98..1912
/gene="Cox-2"
/EC_number="1.14.99.1"
/codon_start=1
/product="cyclooxygenase-2"
/protein_id="AAA58433.1"
/db_xref="GI:181254"
/translation="MLARALLLCAVLALSHTANPCCSHPCQNRGVCMSVGFDQYKCDC
TRTGFYGENCSTPEFLTRIKLFLKPTPNTVHYILTHFKGFWNVVNNIPFLRNAIMSYV
LTSRSHLIDSPPTYNADYGYKSWEAFSNLSYYTRALPPVPDDCPTPLGVKGKKQLPDS
NEIVGKLLLRRKFIPDPQGSNMMFAFFAQHFTHQFFKTDHKRGPAFTNGLGHGVDLNH
IYGETLARQRKLRLFKDGKMKYQIIDGEMYPPTVKDTQAEMIYPPQVPEHLRFAVGQE
VFGLVPGLMMYATIWLREHNRVCDVLKQEHPEWGDEQLFQTSRLILIGETIKIVIEDY
VQHLSGYHFKLKFDPELLFNKQFQYQNRIAAEFNTLYHWHPLLPDTFQIHDQKYNYQQ
FIYNNSILLEHGITQFVESFTRQIAGRVAGGRNVPPAVQKVSQASIDQSRQMKYQSFN
EYRKRFMLKPYESFEELTGEKEMSAELEALYGDIDAVELYPALLVEKPRPDAIFGETM
VEVGAPFSLKGLMGNVICSPAYWKPSTFGGEVGFQIINTASIQSLICNNVKGCPFTSF
SVPDPELIKTVTINASSSRSGLDDINPTVLLKERSTEL"
sig_peptide 98..148
/gene="Cox-2"
mat_peptide 149..1909
/gene="Cox-2"
/product="cyclooxygenase-2"
/EC_number="1.14.99.1"
3'UTR 1913..3387
/gene="Cox-2"
polyA_signal 3369..3374
/gene="Cox-2"
ORIGIN
1 gtccaggaac tcctcagcag cgcctccttc agctccacag ccagacgccc tcagacagca
61 aagcctaccc ccgcgccgcg ccctgccccc cgctgcgatg ctcgccccgc ccctgctgct
121 gtgcgcggtc ctggcgctca gccatacagc aaatccttgc tgttcccacc catgtcaaaa
181 ccgaggtgta tgtatgagtg tgggatttga ccagtataag tgcgattgta cccggacagg
Done
Internet
Start Bioinfo6 - Microsoft Word Bioinfo_course NCBI Sequence Viewe...
EN 21:49

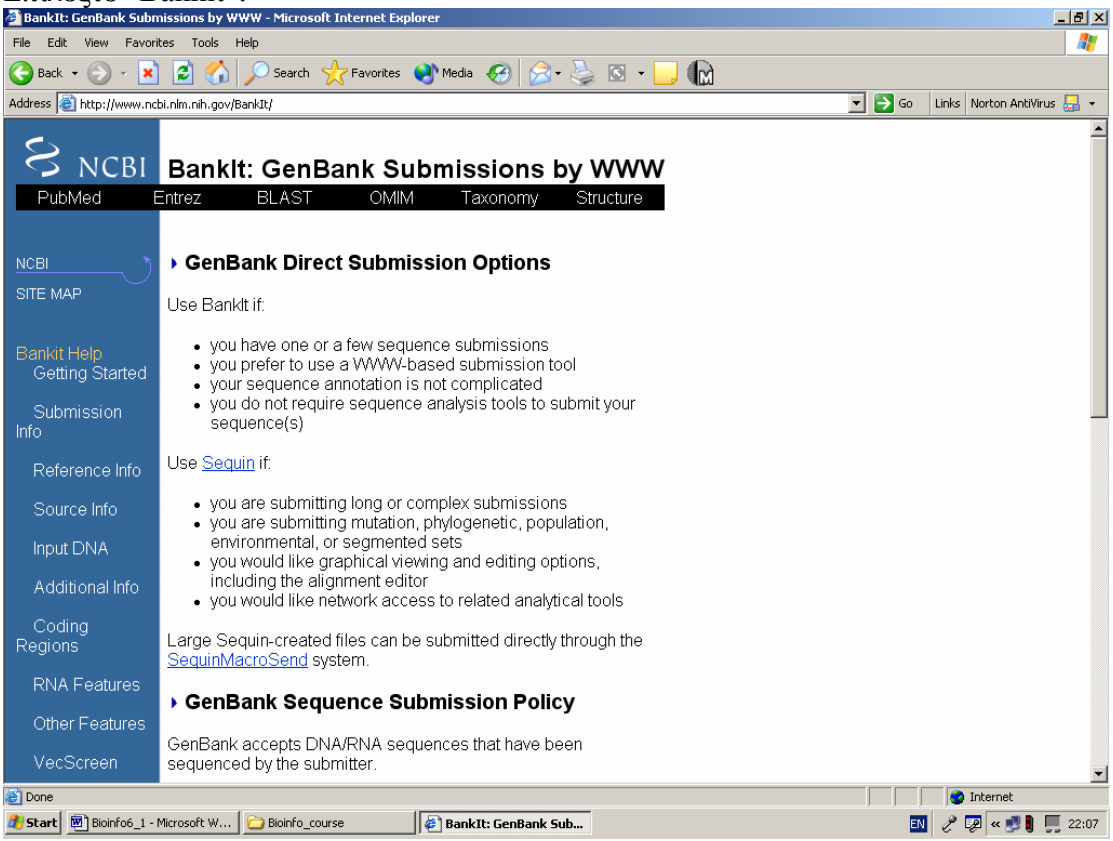

NCBI Sequence Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=181253
Go Links Norton AntiVirus
ORIGIN
1 gtccaggaac tcctcagcag cgcctccttc agctccacag ccagacgccc tcagacagca
61 aagcctaccc ccgcgccgcg ccctgccccc cgctgcgatg ctcgccccgc ccctgctgct
121 gtgcgcggtc ctggcgctca gccatacagc aaatccttgc tgttcccacc catgtcaaaa
181 ccgaggtgta tgtatgagtg tgggatttga ccagtataag tgcgattgta cccggacagg
241 attctatgga gaaaactgct caacaccgga atttttgaca agaataaaat tatttctgaa
301 acccactcca aacacagtgc actacatact tacccacttc aagggatttt ggaacgttgt
361 gaataacatt cccttccttc gaaatgcaat tatgagttat gtgttgacat ccagatcaca
421 tttgattgac agtccaccaa cttacaatgc tgactatggc tacaaaagct gggaagcctt
481 ctctaacctc tcctattata ctagagccct tcctcctgtg cctgatgatt gcccgactcc
541 cttgggtgtc aaaggtaaaa agcagcttcc tgattcaaat gagattgtgg gaaaattgct
601 tctaagaaga aagttcatcc ctgatcccca gggctcaaac atgatgtttg cattctttgc
661 ccagcacttc acgcatcagt ttttcaagac agatcataag cgagggccag ctttcaccaa
721 cgggctgggc catggggtgg acttaaatca tatttacggt gaaactctgg ctagacagcg
781 taaactgcgc cttttcaagg atggaaaaat gaaatatcag ataattgatg gagagatgta
841 tcctcccaca gtcaaagata ctcaggcaga gatgatctac cctcctcaag tccctgagca
901 tctacggttt gctgtggggc aggaggtctt tggtctggtg cctggtctga tgatgtatgc
961 cacaatctgg ctgagggaac acaacagagt atgcgatgtg cttaaacagg agcatcctga
1021 atggggtgat gagcagttgt tccagacaag caggctaata ctgataggag agactattaa
1081 gattgtgatt gaagattatg tgcaacactt gagtggctat cacttcaaac tgaaatttga
1141 cccagaacta cttttcaaca aacaattcca gtaccaaaat cgtattgctg ctgaatttaa
1201 caccctctat cactggcatc cccttctgcc tgacaccttt caaattcatg accagaaata
1261 caactatcaa cagtttatct acaacaactc tatattgctg gaacatggaa ttacccagtt
1321 tgttgaatca ttcaccaggc aaattgctgg cagggttgct ggtggtagga atgttccacc
1381 cgcagtacag aaagtatcac aggcttccat tgaccagagc aggcagatga aataccagtc
1441 ttttaatgag taccgcaaac gctttatgct gaagccctat gaatcatttg aagaacttac
1501 aggagaaaag gaaatgtctg cagagttgga agcactctat ggtgacatcg atgctgtgga
1561 gctgtatcct gcccttctgg tagaaaagcc tcggccagat gccatctttg gtgaaaccat
1621 ggtagaagtt ggagcaccat tctccttgaa aggacttatg ggtaatgtta tatgttctcc
1681 tgcctactgg aagccaagca cttttggtgg agaagtgggt tttcaaatca tcaacactgc
1741 ctcaattcag tctctcatct gcaataacgt gaagggctgt ccctttactt cattcagtgt
1801 tccagatcca gagctcatta aaacagtcac catcaatgca agttcttccc gctccggact
1861 agatgatatc aatcccacag tactactaaa agaacgttcg actgaactgt agaagtctaa
1921 tgatcatatt tatttattta tatgaaccat gtctattaat ttaattattt aataatattt
1981 atattaaact ccttatgtta cttaacatct tctgtaacag aagtcagtac tcctgttgcg
2041 gagaaaggag tcatacttgt gaagacttttt atgtcactac tctaaagatt ttgctgttgc
2101 tgttaagttt ggaaaacagt ttttattctg ttttataaac cagagagaaa tgagttttga
2161 cgtctttta cttgaatttc aacttatatt ataaggacga aagtaaagat gtttgaatac
Done
Internet
Start Bioinfo6 - Microsoft Word Bioinfo_course NCBI Sequence Viewe...
EN 21:50



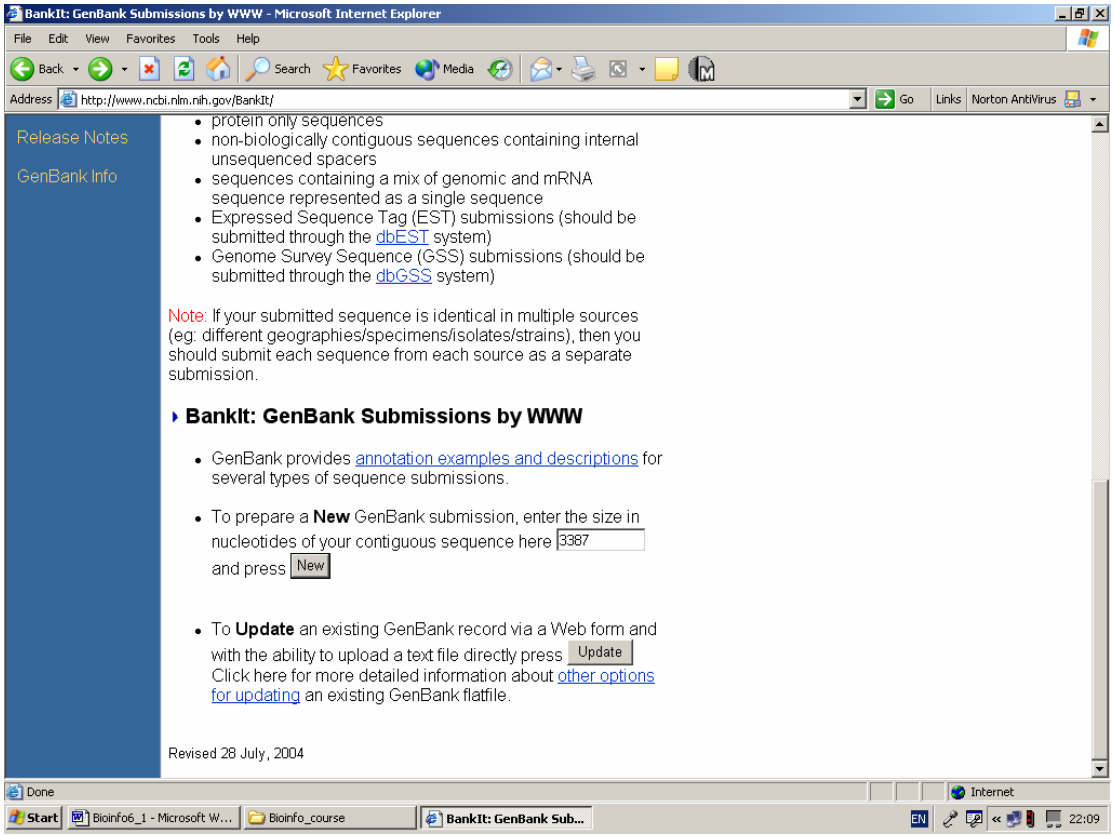
Για να καταχωρήσετε μια νέα ακολουθία ή να ενημερώσετε μια ήδη υπάρχουσα, επιλέξετε “Submit to GenBank” στην ιστοσελίδα παρακάτω. Στην συνέχεια ακολουθήσετε τις οδηγίες που εμφανίζονται. Το “Banklt” είναι ένα εργαλείο που μπορείτε να χρησιμοποιήσετε για να καταχωρήσετε εύκολα και απλά μια νέα αλληλουχία. Η δομή της καταχώρησης θα πρέπει να μοιάζει με αυτή της HUMCYCLOX.



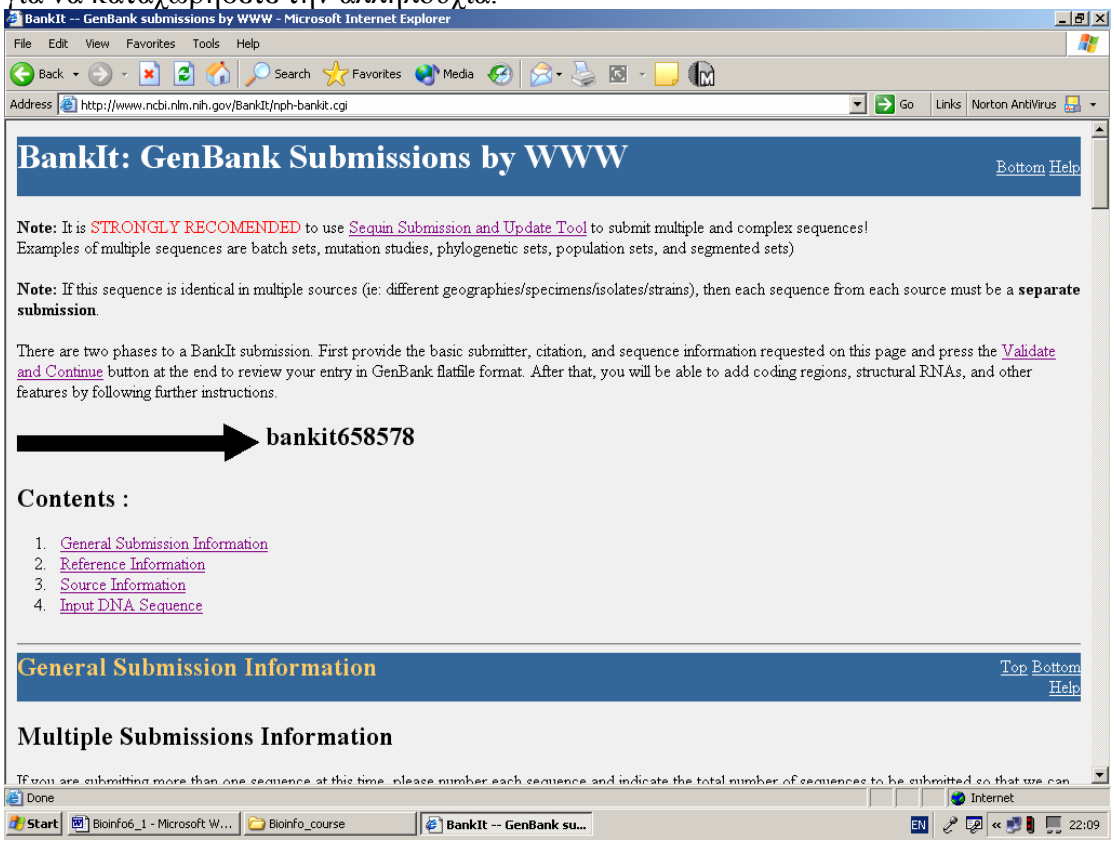
Επιλέξτε “BankIt”.



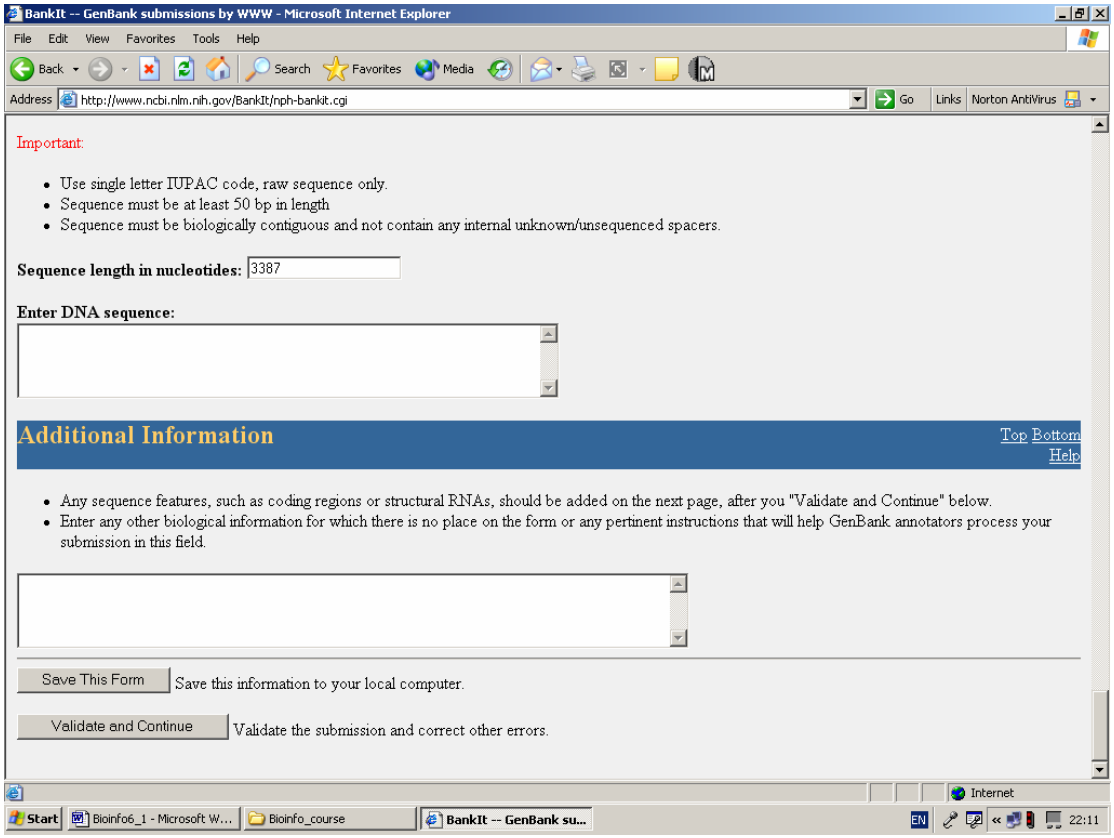
Καταχωρίστε το μέγεθος της αλληλουχίας, π.χ. 3387, και επιλέξτε “New”. Για ενημέρωση επιλέξτε “Update”.



Κυλήστε την οθόνη προς τα κάτω για να δείτε όλες τις πληροφορίες που χρειάζονται για να καταχωρήσετε την αλληλουχία.



Αφού έχετε συμπληρώσει όλα τα απαιτούμενα πεδία στην ιστοσελίδα, πληκτρολογήστε στο πεδίο παρακάτω την αλληλουχία, και επιλέξτε “Validate and Continue”.



Ζευγαρωτή αντιστοιχία αλληλουχιών

Αναζήτηση στις βάσεις δεδομένων

Η αναζήτηση όμοιων αλληλουχιών σε βάσεις δεδομένων μας δίνει τη δυνατότητα ανάκτησης αλληλουχιών, που είναι όμοιες με μια ζητούμενη (query) αλληλουχία, και επίσης τη δυνατότητα ποσοτικοποίησης αυτής της ομοιότητας. Το μέγεθος της ομοιότητας επιτρέπει την αναγνώριση της δομής, της λειτουργίας, ή της οικογενείας της ζητούμενης αλληλουχίας.
Δύο αλληλουχίες DNA ή αλληλουχίες πρωτεϊνών που είναι πολύ όμοιες πιθανόν να έχουν σχετιζόμενες λειτουργίες και επίσης μπορεί να σχετίζονται επειδή έχουν έναν κοινό πρόγονο.

Αντιστοιχία αλληλουχιών

Μια από τις πιο χρήσιμες αναπαραστάσεις της ομοιότητας αλληλουχιών είναι η αντιστοιχία. Ας θεωρήσουμε ένα απλό παράδειγμα όπου θέλουμε να συγκρίνουμε τις δύο παρακάτω αλληλουχίες DNA:

X= A A T C T G A T A G A A G C C C T A
Y= C C A A T C C A G A A C G C C C A

Μπορούμε να μετασχηματίσουμε την X σε Y (ή αντίστροφα) με μια σειρά απλών αλλαγών βάσεων, μεταλλάξεων ή επεμβατικών λειτουργιών. Οι επιτρεπτές λειτουργίες είναι:

- Ομοιότητα (match): παραμένει η βάση αμετάβλητη
- Μη- ομοιότητα (mismatch): αντικατάσταση μιας βάσης από διαφορετική βάση
- Κενό (gap): εισαγωγή / διαγραφή μιας βάσης

Μια αντιστοιχία της X και Y είναι η απεικόνιση των επεμβατικών λειτουργιών, οι οποίες είναι απαραίτητες για τον μετασχηματισμό μιας σειράς σε μια άλλη.
Υπάρχει μεγάλος αριθμός πιθανών αντιστοιχιών της X κα Y, που αντιστοιχούν σε όλους τους δυνατούς συνδυασμούς όπου οι αλληλουχίες θα μπορούσαν να αποκλίνουν από μια κοινή προγονική αλληλουχία. Μια τέτοια αντιστοιχία είναι η παρακάτω:

X		**-**	**-**	**A**	**A**	**T**	**C**	**T**	**G**	**A**	**T**		**A**	**G**	**A**	**A**	**G**	**C**	**C**	**C**	**T**	**A**
				:	**:**	**:**	**:**		**:**	**:**	*****	**:**			**:**	*****	**:**	**:**	**:**	**:**		**:**
Y		**C**	**C**	**A**	**A**	**T**	**C**	**-**	**G**	**A**	**G**		**A**	**-**	**A**	**C**	**G**	**C**	**C**	**C**	**-**	**A**
Θέση		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	

Όπου,
: σημαίνει ομοιότητα
***** σημαίνει μη-ομοιότητα
- σημαίνει κενό λόγω της εισαγωγής μιας βάσης σε μια αλληλουχία, ή αντίστοιχα η διαγραφή μιας βάσης στην άλλη αλληλουχία.

Σύμφωνα με την παραπάνω αντιστοιχία, για τον μετασχηματισμό της X στην Y θα πρέπει να γίνει:

Αντικατάσταση της G από T στη θέση 10

Αντικατάσταση της A από C στη θέση 14
Εισαγωγή της C στις θέσεις 1, 2
Διαγραφή της T στις θέσεις 7, 19
Διαγραφή της G στην θέση 12

Οπότε η αντιστοιχία περιέχει 13 ομοιότητες, 2 μη-ομοιότητες και 5 κενά.
Το συνολικό μήκος της αντιστοιχίας είναι 20.

Ένα μέτρο της ομοιότητας σειράς είναι το ποσοστό ομολογίας, το οποίο ορίζεται ως το ποσοστό των ομοιοτήτων στο πλήρες μήκος της αντιστοιχίας, όπου σε αυτή την περίπτωση είναι: (13/20)x100=65%.
Το ποσοστό της ομολογίας ορίζεται κάποιες φορές ως το ποσοστό των ομοιοτήτων που περιέχονται μέσα στο μήκος της μικρότερης αλληλουχίας.
Για οποιοδήποτε ζεύγος αλληλουχιών θα υπάρχουν πολλαπλές δυνατές αντιστοιχίες. Για παράδειγμα, χρησιμοποιώντας μερικές διαφορετικές επεμβατικές λειτουργίες, μια εναλλακτική αντιστοιχία για τις παραπάνω σειρές X και Y είναι η παρακάτω:

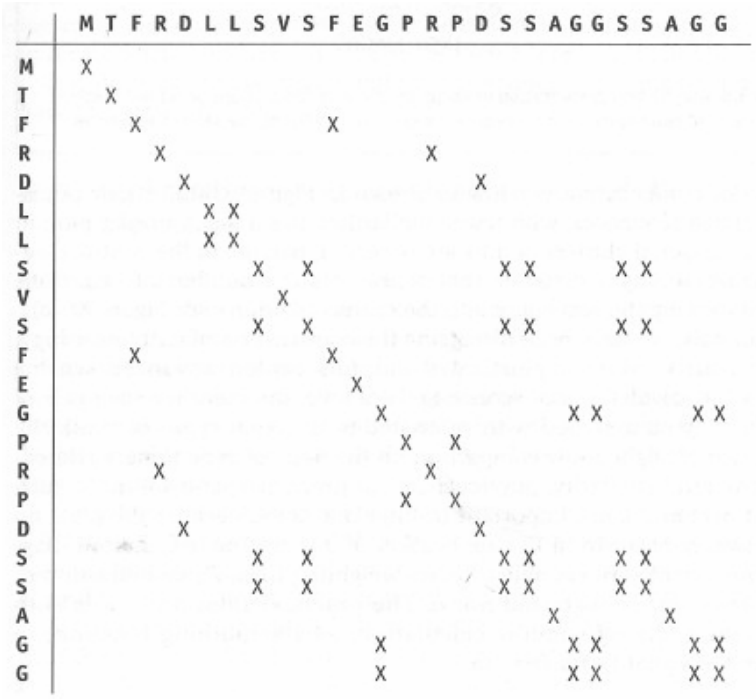
```
X        - - A A T C T G A T A G A A - G C C C T A
                 :   :   :   :     :       :   :   :   :       :   :   :   :     :
Y        C C A A T C - G - - A G A A C G C C C - A
Θέση     1 2  3  4 5  6 7 8  9 10 11 12 13 14 15 16 17 18 19 20 21
```

Η οποία περιέχει 14 ομοιότητες, 0 αντικαταστάσεις και 7 κενά. Τώρα το μήκος της αντιστοιχίας είναι 21 και το ποσοστό ομολογίας έχει αυξηθεί σε (14/21)x100=66.7%.

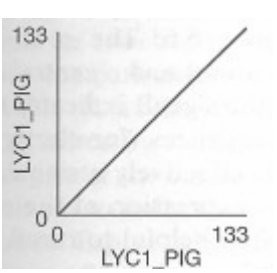
Στατιστικές μετρήσεις για τη σημαντικότητα αντιστοιχίας στην αναζήτηση σε βάσεις δεδομένων
Η αντιστοιχία αλληλουχιών πραγματοποιείται με τη χρήση προγραμμάτων υπολογιστών. Αυτά τα προγράμματα παρέχουν κάποια στατιστική εκτίμηση δηλώνοντας το επίπεδο αξιοπιστίας που θα πρέπει να σχετίζεται σε μια αντιστοιχία. Τα συνηθισμένα στατιστικά μεγέθη είναι το p-value και E-value. Το p-value σχετίζει το αποτέλεσμα μιας αντιστοιχίας με την πιθανότητα να είναι τυχαίο (όσο πιο πολύ προσεγγίζει το μηδέν, τόσο μεγαλύτερη αξιοπιστία υπάρχει ότι το αποτέλεσμα είναι πραγματικό). Το E-value περιγράφει τον αριθμό επιτυχιών (ομοιοτήτων) που αναμένεται να είναι τυχαία στην αναζήτηση μιας βάσης δεδομένων συγκεκριμένου μεγέθους (όταν το E-value πάρει την τιμή 1 για ένα ταίριασμα, αυτό μπορεί να ερμηνευτεί ότι στην τρέχουσα έρευνα, αναμένεται μόνο από τύχη να βρεθεί μια ομοιότητα με ίδιο αποτέλεσμα. Μια τιμή 0 δηλώνει ότι κανένα δεν αναμένεται να είναι τυχαίο, δηλ. είναι απίθανο η αντιστοιχία να είναι από τυχαία ομοιότητα).

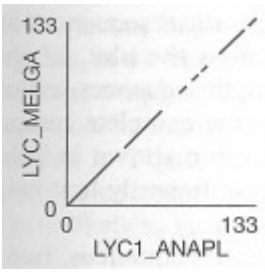
Διάγραμμα ακίδων

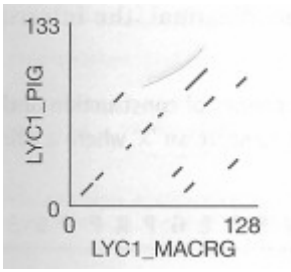
Το διάγραμμα ακίδων είναι μια γραφική παράσταση της ομοιότητας δύο αλληλουχιών. Ας θεωρήσουμε δύο σειρές, A και B, με διαφορετικά μήκη. Σε ένα διάγραμμα ακίδων, δημιουργούμε έναν ορθογώνιο πίνακα όπου π.χ. τα αμινοξέα (residues) της A τοποθετούνται πάνω στο x-άξονα και τα αμινοξέα του B πάνω στο y-άξονα. Τα κελιά για τα οποία ισχύει Ai=Bj παίρνουν την τιμή 1 αλλιώς την τιμή 0.
Στην γραφική απεικόνιση το 1 συμβολίζεται με ακίδα και το 0 με κενό και ο πίνακας παρουσιάζεται με ένα διάγραμμα. Για παράδειγμα, από τη σύγκριση των δύο σειρών δημιουργήθηκε το παρακάτω διάγραμμα.



Το διάγραμμα χαρακτηρίζεται από μερικές τυχαίες ακίδες (τυχαίο σφάλμα/θόρυβος) και μια κεντρική διαγώνια γραμμή (μονοπάτι), όπου οι συνεχόμενες ακίδες υψηλής πυκνότητας (σήμα) δηλώνουν τις περιοχές με την μέγιστη ομοιότητα μεταξύ των δύο αλληλουχιών.







Δύο πανομοιότυπες αλληλουχίες απεικονίζονται με μία απλή συνεχόμενη διαγώνια γραμμή κατά μήκος του διαγράμματος. Δύο παρόμοιες αλληλουχίες θα απεικονίζονται με μια διακεκομμένη διαγώνια γραμμή, όπου οι περιοχές με τις διακοπές δηλώνουν μη-ομοιότητα. Ενώ, δύο διαφορετικές αλλά σχετιζόμενες αλληλουχίες θα απεικονίζονται από διαγώνιες ομάδες ακίδων, παράλληλες με την κεντρική διαγώνιο.

Δυναμικός προγραμματισμός – Αλγόριθμος του Needleman και Wunsch

Στην αντιστοιχία σειρών, η καλύτερη αντιστοιχία, ή το καλύτερο μονοπάτι, μπορεί να βρεθεί χρησιμοποιώντας δυναμικό προγραμματισμό.

Ο δυναμικός προγραμματισμός είναι μια τεχνική βελτιστοποίησης για την ανάλυση βαθμολογημένων πινάκων, ο οποίος βρίσκει το υψηλότερα βαθμολογημένο μονοπάτι σε ένα πίνακα, δηλ. βρίσκει την καλύτερη αντιστοιχία μεταξύ δύο σειρών.

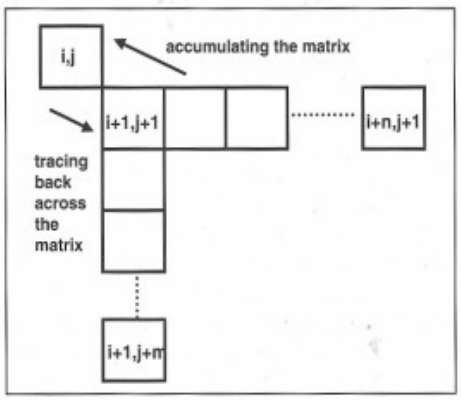
Μεταξύ δύο σειρών που φαίνονται διαφορετικές, συχνά υπάρχουν πολλαπλές δυνατές αντιστοιχίες. Ο δυναμικός προγραμματισμός επιτρέπει τον έλεγχο διαφορετικών μονοπατιών που αντιστοιχούν σε διαφορετικές αντιστοιχίες με υψηλή ομολογία (βαθμολόγηση),

συνυπολογίζοντας διάφορες παραμέτρους (π.χ. κυρώσεις από κενά). Στην ουσία, προσπαθεί να ταιριάξει τον μέγιστο αριθμό από ζεύγη πανομοιότυπων αμινοξέων αλλά και ταυτόχρονα επιτρέποντας το ελάχιστο αριθμό εισαγωγών και διαγραφών σε δύο σειρές,
Στο τέλος, επιλέγεται το καλύτερο από όλα τα μονοπάτια, ως η τελική αντιστοιχία.

Στο δυναμικό προγραμματισμό περιλαμβάνονται τρία βασικά βήματα:

- Βαθμολογία του δισδιάστατου (2D) βαθμολογημένου πίνακα ανάλογα με τις ομοιότητες των αμινοξέων που συγκρίνονται. Ο πίνακας βαθμολογείται με την τιμή 1 για το πανομοιότυπο ζεύγος αμινοξέων και την τιμή 0 για νη-ομοιότητα.
- Συγκέντρωση των βαθμολογιών στον 2D πίνακα, μειώνοντας την βαθμολογία για τις εισαγωγές/διαγραφές. Οι βαθμολογίες στον πίνακα αθροίζονται από την κάτω δεξιά γωνία του πίνακα μέχρι την πάνω αριστερή γωνιά του πίνακα. Αυτό γίνεται με τον υπολογισμό των βαθμολογιών σε μια γραμμή την φορά, ξεκινώντας από το πιο δεξί κελί της συγκεκριμένης γραμμής. Μετά για κάθε κελί (i,j) στη γραμμή, η καλύτερη βαθμολογία από όλα τα πιθανά μονοπάτια που οδηγούνε σε αυτό το κελί αθροίζεται στην βαθμολογία που ήδη υπάρχει στο συγκεκριμένο κελί. Η άθροιση των βαθμολογιών σε κάθε θέση στον πίνακα εκφράζεται μαθηματικά με τον παρακάτω τύπο:
  $S_{i,j} = S_{i,j} + \max\{S_{i+1,j+1} , S_{i+m,j+1} - g , S_{i+1,j+m} - g\}$, όπου $S_{i+m,j+1}$ είναι η μέγιστη παρατηρούμενη βαθμολογία στη γραμμή j+1, $S_{i+1,j+m}$ είναι η μέγιστη παρατηρούμενη βαθμολογία στη στήλη i+1 και το g είναι η κύρωση αν εισαχθεί ένα κενό.
- Προσδιορισμός του μονοπατιού με τη μεγαλύτερη βαθμολογία στον πίνακα. Το μονοπάτι θα ξεκινήσει από το κελί με τη μεγαλύτερη βαθμολογία στη γραμμή της κορυφής ή στην πιο αριστερά στήλη του πίνακα και θα καθοριστεί από εκεί με παρόμοιο τρόπο.

Το καλύτερο μονοπάτι από κάθε (i,j) κελί θα προκύψει από τη μεγαλύτερη βαθμολογία των κελιών, της γραμμής (j+1) ή της στήλης (i+1) που βρίσκεται από δεξιά και προς τα κάτω του, στο πίνακα.



	A	D	L	G	A	V	F	A	L	C	D	R	Y	F	Q
A	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0
D	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
L	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
G	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
D	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

	A	D	L	G	A	V	F	A	L	C	D	R	Y	F	Q
A	1	0	0	0	1	0	0	1	0	4	3	2	1	1	0
D	0	1	0	0	0	0	0	0	0	4	4	2	1	1	0
L	0	0	1	0	0	0	0	0	1	4	3	2	1	1	0
G	0	0	0	1	0	0	0	0	5	4	3	2	1	1	0
R	0	0	0	0	0	0	0	0	5	4	3	3	1	1	0
T	0	0	0	0	0	0	0	0	5	4	3	2	1	1	0
Q	0	0	0	0	0	0	0	0	5	4	3	2	1	1	1
N	0	0	0	0	0	0	0	0	5	4	3	2	1	1	0
C	0	0	0	0	0	0	0	0	4	5	3	2	1	1	0
D	0	1	0	0	0	0	0	0	3	3	4	2	1	1	0
R	0	0	0	0	0	0	0	0	2	2	2	3	1	1	0
Y	0	0	0	0	0	0	0	0	2	2	2	2	2	1	0
Y	0	0	0	0	0	0	0	0	1	1	1	1	2	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

	A	D	L	G	A	V	F	A	L	C	D	R	Y	F	Q
A	9	7	6	6	7	6	6	7	5	4	3	2	1	1	0
D	7	8	6	6	6	6	6	6	5	4	4	2	1	1	0
L	6	6	7	5	5	5	5	5	6	4	3	2	1	1	0
G	5	5	5	6	5	5	5	5	5	4	3	2	1	1	0
R	5	5	5	5	5	5	5	5	5	4	3	3	1	1	0
T	5	5	5	5	5	5	5	5	5	4	3	2	1	1	0
Q	5	5	5	5	5	5	5	5	5	4	3	2	1	1	1
N	5	5	5	5	5	5	5	5	5	4	3	2	1	1	0
C	4	4	4	4	4	4	4	4	4	5	3	2	1	1	0
D	3	4	3	3	3	3	3	3	3	3	4	2	1	1	0
R	2	2	2	2	2	2	2	2	2	2	2	3	1	1	0
Y	2	2	2	2	2	2	2	2	2	2	2	2	2	1	0
Y	1	1	1	1	1	1	1	1	1	1	1	1	2	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Η τελική αντιστοιχία είναι

ADLGAVFALCDRYFQ
: : : : : : : : :
ADLGRTQN-CDR YYQ

Προχωρημένα στοιχεία για τη Σύγκριση Αλληλουχιών (Σειρών)

Για ποιο λόγο ασχολούμαστε με τη σύγκριση σειρών;

Το DNA (Deoxyribonucleic acid), το RNA (Ribonucleic acid) και οι πρωτεΐνες αποτελούν τα βασικά βιομόρια για όλους τους ζώντες οργανισμούς. Σε πολύ γενικές γραμμές, βασικός ρόλος του DNA είναι η αποθήκευση πληροφοριών, του RNA η κατανομή των πληροφοριών (το RNA είναι αναλώσιμο), και οι πρωτεΐνες ενεργούν ως λειτουργικοί πράκτορες, οι οποίοι παίζουν ζωτικό ρόλο στον έλεγχο, την επιρροή και τη μεταβολή των κυτταρικών λειτουργιών και της φαινοτυπικής συμπεριφοράς.

Χάριν απλούστευσης μπορούμε να σκεφτούμε τα μόρια αυτά ως γραμμικές γραμματοσειρές, οι οποίες αποτελούνται από συγκεκριμένες δομικές μονάδες. Στην περίπτωση του DNA και του RNA οι δομικές αυτές μονάδες ονομάζονται νουκλεοτίδια (nucleotides) ή βάσεις (bases) και είναι οι εξής : αδενίνη (adenine, A), κυτοσίνη (cytosine, C). η γουανίνη (guanine, G) και η θυμίνη (thymine, T). Οι τρεις πρώτες βάσεις είναι κοινές για το DNA και το RNA, με μόνη διαφορά ότι στο RNA η θυμίνη αντικαθίσταται από την ουρακίλη (uracil, U). Στην περίπτωση των πρωτεϊνών, οι δομικές μονάδες είναι 20 και ονομάζονται αμινοξέα (amino acids), ή **amino acid residues**, τα οποία σε κατώτερο επίπεδο απαρτίζονται από τις ίδιες τέσσερις βάσεις που συγκροτούν και την αλυσίδα του DNA (αυτό γίνεται εύκολα αντιληπτό αρκεί να αναλογιστούμε ότι η σύνθεση των πρωτεϊνών γίνεται από το DNA με τη βοήθεια του RNA).

Κατά τα τελευταία χρόνια μέσα από διάφορα projects έχει γίνει γνωστός ένας μεγάλος αριθμός σειρών. Ενδεικτικό είναι ότι το 1989, το sequencing ενός γονιδίου 1.8kb αποτελούσε θέμα διδακτορικής διατριβής. Το 1993 το ίδιο θέμα ήταν προπτυχιακό θέμα, ενώ το 2000 σε μη-γενετικό εργαστήριο κατασκευάζονταν 40kb σειράς. Εφόσον πλέον υπάρχει διαθέσιμος ένας μεγάλος αριθμός σειρών, το επόμενο βήμα είναι να γίνει γνωστή και η λειτουργία που εκφράζει καθεμία από αυτές τις σειρές.

Σύγκριση σειρών πρωτεϊνών vs. Σύγκριση τριτοταγών δομών πρωτεϊνών

Η σειρά των αμινοξέων που απαρτίζουν μία πρωτεΐνη δίνει σημαντικά στοιχεία για τη λειτουργία της. Και η διαμόρφωση όμως μίας πρωτείνης στο χώρο καθορίζει σε πολύ μεγάλο βαθμό τη λειτουργία της. Επιπλέον, μέσα από την εξέλιξη η τρισδιάστατη δομή της πρωτείνης διατηρείται καλύτερα απ' ό,τι η σειρά των αμινοξέων της, δίνοντας έτσι περισσότερες πληροφορίες σχετικά με την εξελικτική της πορεία. Επομένως, ένα φυσικό ερώτημα είναι για ποιο λόγο να προτιμηθεί η σύγκριση των σειρών δύο πρωτεϊνών έναντι στη σύγκριση των τρισδιάστατων δομών αυτών, η οποία και απλούστερη μοιάζει και πιο αποδοτική. Ο λόγος είναι ότι ο αριθμός των γνωστών σήμερα σειρών πρωτεϊνών είναι περίπου τριάντα φορές μεγαλύτερος από τον αριθμό των τριτοταγών δομών και επιπλέον οι διαθέσιμες μέθοδοι σύγκρισης σειρών είναι σε γενικές γραμμές πολύ ταχύτερες.

Ομολογία σειρών (Sequence homology)

Γνωρίζοντας ήδη μία σειρά βάσεων ή αμινοξέων, αυτό που πρέπει να κάνουμε προκειμένου να προσδιορίζουμε τη λειτουργία της είναι να βρούμε μία ομόλογη αυτής σειρά, για την οποία να είναι γνωστή η λειτουργία της. Ως ομόλογες (**homologs**) χαρακτηρίζονται σειρές ή δομές οι οποίες έχουν

προέλθει από ένα κοινό πρόγονο μέσα από την εξελικτική διαφοροποίηση. Κατ' αυτόν τον τρόπο δύο ομόλογες σειρές διατηρούν κοινά χαρακτηριστικά, χωρίς όμως να είναι πανομοιότυπες, ενώ γνώση της λειτουργίας της μίας αποτελεί πολύτιμο οδηγό στον καθορισμό της λειτουργίας της άλλης. Η ομολογία (**homology**) δεν μπορεί να προσδιοριστεί άμεσα, αλλά πρέπει να διαπιστωθεί μέσω της ομοιότητας των εν λόγω σειρών και των δομών.

Το πρώτο σημείο στο οποίο πρέπει να σταθούμε κατά τη σύγκριση δύο σειρών είναι το αν οι σειρές αυτές συσχετίζονται. Αυτό συνήθως επιτυγχάνεται αρχικά με τη σύγκριση / αντιπαράθεση (**aligning**) των δύο σειρών, ή τμημάτων αυτών και εν συνεχεία με την απόφαση εάν η όποια σχέση προκύψει οφείλεται επειδή όντως οι σειρές σχετίζονται ή είναι απλά τυχαίο γεγονός. Για το διαχωρισμό των δύο αυτών περιπτώσεων, καθοριστικής σημασίας είναι το **scoring scheme** το οποίο θα χρησιμοποιηθεί για την αξιολόγηση της συσχέτισης.

Scoring model

Οι αλλαγές οι οποίες παρουσιάζονται ανάμεσα σε δύο ομόλογες σειρές κατά την πορεία της εξέλιξης, οφείλονται σε μεταλλάξεις (**mutations**) και στην επιλογή (**selection**). Όσον αφορά τις μεταλλάξεις, οι βασικές τους μορφές είναι οι αντικαταστάσεις (**substitutions**) βάσεων στην περίπτωση σειράς DNA ή RNA, ή αμινοξέων (residues) σε μία σειρά πρωτεΐνης, με κάποια άλλη βάση ή με αμινοξύ αντιστοίχως. Επίσης, οι εισαγωγές (**insertions**) και οι διαγραφές (**deletions**), οι οποίες αφορούν αντιστοίχως την προσθήκη ή τη διαγραφή / παράλειψη μίας βάσης ή ενός αμινοξέως από τη σειρά. Οι εισαγωγές και οι διαγραφές αναφέρονται με έναν όρο ως "**κενά**" (**gaps**) ή και ως "**indels**". Αξίζει να αναφέρουμε ότι ορισμένες μορφές μεταλλάξεων εμφανίζονται πιο συχνά απ' ό,τι άλλες, καθώς και ότι κάθε τμήμα της σειράς, ανάλογα με τη μορφολογία του, ευνοεί περισσότερο κάποια μορφή μετάλλαξης έναντι των υπολοίπων. Έτσι, για παράδειγμα, στις πρωτεΐνες, είναι σύνηθες οι εισαγωγές να παρουσιάζονται σε loops τα οποία συνδέουν τις δευτεροταγείς δομές (α-helices, β-sheets) , όπου και υπάρχει μικρότερη πιθανότητα να διαταράξουν τους υπάρχοντες δεσμούς υδρογόνου.

Ένας τρόπος για να ορίσουμε την ολική βαθμολογία την οποία θα αποδώσουμε στην εκάστοτε αντιστοίχηση σειρών είναι ως το άθροισμα των όρων (επιμέρους βαθμών) που αποδίδονται για κάθε αντιστοίχηση βάσεων / αμινοξέων μεταξύ των δύο σειρών, συν τους όρους (επιμέρους βαθμούς) που αποδίδονται για κάθε κενό στις σειρές. Είναι αναμενόμενο οι ομοιότητες και οι συντηρητικές αντικαταστάσεις (όπου με τον όρο αυτόν εννοούμε αντικαταστάσεις βάσεων / αμινοξέων από συγγενικές τους βάσεις / συγγενικά αμινοξέα, έτσι ώστε να μην επηρεάζεται η λειτουργία της σειράς) να είναι πιο συνήθεις σε μία αντιστοίχηση σειρών, επομένως σε αυτές αποδίδουμε θετική επιμέρους βαθμολογία. Αντιθέτως, οι μη συντηρητικές αλλαγές, στις οποίες οφείλονται και οι σημαντικές διαφοροποιήσεις στη λειτουργία, δεν είναι τόσο κοινές, επομένως σε αυτές αποδίδουμε αρνητικούς βαθμούς. Με τον τρόπο αυτό επιτυγχάνεται η τελική βαθμολογία της αντιστοιχίας να είναι θετικός αριθμός.

Η χρήση ενός αθροιστικού τρόπου βαθμολόγησης όπως περιγράφηκε, προϋποθέτει ότι δεν υπάρχει εξάρτηση ανάμεσα στις τυχόν μεταλλάξεις που παρουσιάζονται σε διαφορετικές βάσεις / διαφορετικά αμινοξέα μίας σειράς. Αυτή η παραδοχή αποδεικνύεται ευσταθής για την περίπτωση που γίνεται αντιστοίχηση σειρών DNA ή πρωτεϊνών, έστω κι αν είναι γνωστό ότι οι αλληλεπιδράσεις μεταξύ των αμινοξέων παίζουν πολύ σημαντικό ρόλο στη διαμόρφωση της τριτοταγούς δομής μίας

πρωτεΐνης. Στην περίπτωση όμως που οι υπό μελέτη σειρές είναι σειρές RNA δεν μπορούμε να αγνοήσουμε τη σχέση ανάμεσα στις επιμέρους μεταλλάξεις της σειράς. Εδώ δε θα ασχοληθούμε όμως περαιτέρω με σειρές RNA, αλλά θα επικεντρωθούμε σε σειρές πρωτεϊνών, έχοντας όμως υπ' όψην ότι όλα όσα αναφέρουμε είναι εφαρμόσιμα και σε σειρές DNA.

Παράδειγμα:

Έστω ότι έχουμε δύο σειρές και θέλουμε να βρούμε τη μεταξύ τους σχέση. Τα βήματα που θα ακολουθήσουμε σε γενικές γραμμές είναι τα ακόλουθα:

- 1ο βήμα:

  Εντοπίζουμε τις ακριβείς αντιστοιχίες μεταξύ των δύο σειρών και αποδίδουμε βαθμολογία στην καθεμία.

  ```
  ACCGGTATCC---GAC
  |||   ||||  ||||||
  ACC--TATCTTAGGAC
  ```

- 2ο βήμα:

  Εντοπίζουμε τις συντηρητικές αντικαταστάσεις και αποδίδουμε σε αυτές τους ανάλογους βαθμούς.

  ```
  ACCGGTATCC---GAC
  |||   ||||  ||||||
  ACC--TATCTTAGGAC
  ```

- 3ο βήμα:

  Αποδίδουμε την κατάλληλη βαθμολογία (ή ποινή) σε κάθε κενό ή εισαγωγή στις σειρές. Το μήκος ενός κενού είναι ο αριθμός των indels που το αποτελούν. Στο απλό αυτό παράδειγμα συναντούμε δύο κενά (ένα κενό σε κάθε σειρά), μήκους 2 και 3 indels.

  ```
  ACCGGTATCC---GAC
  |||   ||||  ||||||
  ACC--TATCTTAGGAC
  ```

Βαθμολόγηση / Ποινές κενών (Gap penalties)

Η επιλογή του τρόπου βαθμολόγησης των κενών σε μία σειρά είναι μία από τις σημαντικότερες αποφάσεις που πρέπει να ληφθούν στη διαδικασία αντιστοίχησης δύο σειρών. Η επιβολή ποινής για κάθε κενό που εισάγεται σε μία αντιστοίχηση κρίνεται αναγκαία, καθώς τα κενά αυξάνουν την αβεβαιότητα στην αντιστοίχηση και βοηθούν στην απόφαση εάν είναι προτιμότερο να εισαχθεί ένα κενό στην αντιστοίχηση ή είναι πιο 'αποδοτικό' (θα αποδόσει μεγαλύτερη βαθμολογία στην αντοστοίχηση) εάν γίνει αντιστοίχηση μεταξύ δύο ανόμοιων residues. Από βιολογικής απόψεως θεωρείται πιο έυκολο για μία πρωτεϊνη να 'δεχθεί' την αντικατάσταση ενός residue σε μία θέση, αντί για την εισαγωγή ή διαγραφή τμημάτων της αλληλουχίας. Επομένως τα κενά (gaps)/ εισαγωγές

(insertions) θα έπρεπε να είναι πιο σπάνια από τις αντικαταστάσεις. Εάν τα κενά εισάγονταν χωρίς επιβολή κατάλληλης ποινής, τότε θα μπορούσαν να εισάγονται αυθαίρετα και τελικά να προκύπτει αντιστοιχία μεταξύ οποιονδήποτε σειρών, ακόμα και μεταξύ σειρών τελείως άσχετων μεταξύ τους. Υπάρχουν ορισμένα σημεία τα οποία θα πρέπει να ληφθούν υπόψην ανάλογα με την εκάστοτε περίπτωση, όπως σε περίπτωση που το gap penalty είναι χαμηλό, αυτό αυτομάτως επιτρέπει να θεωρηθεί επιτυχής (αφού η βαθμολόγηση που θα της αποδοθεί θα είναι υψηλή) ακόμα και η αντιστοίχηση ανάμεσα σε δύο μη σχετιζόμενες ή τυχαίες αλληλουχίες. Επίσης, για παράδειγμα, μπορεί να γίνει διαχωρισμός ως προς τη βαθμολόγηση των introns έναντι των exons, ή των κενών που συναντώνται σε περιοχές κωδικοποίησης μίας πρωτεΐνης, κ.ο.κ.

Ενδεικτικά να αναφέρουμε ότι ένας τρόπος για την βαθμολόγηση των κενών θα ήταν να τεθεί ως συνθήκη ότι κάθε νέο κενό που θα συναντάται στη σειρά θα προσδίδει μικρότερη ποινή στην ολική βαθμολογία (ποινή ανοιχτού κενού – gap open penalty). Αποτέλεσμα αυτού θα είναι τα αρχικά κενά θα θεωρούνται πιο σημαντικά από αυτά που έπονται (ποινή κενού επέκτασης – gap extension penalty). Επίσης, μία ακόμα προσέγγιση θα ήταν να χρησιμοποιηθεί μία αυθαίρετη συνάρτηση βασισμένη στο μήκος κάθε κενού.

Εδώ δε θα εμβαθύνουμε στον τρόπο βαθμολόγησης κενών, παρά θα χρησιμοποιήσουμε μία εξαιρετικά απλή, αλλά και αποτελεσματική προσέγγιση: η βαθμολογία / ποινή θα είναι γραμμικός ανάλογη με το μήκος του εκάστοτε κενού . Παραδείγματος χάριν, μία συνάρτηση της μορφής

$$\gamma(g) = -gd,$$

όπου g είναι το μήκος του κενού, d είναι μία σταθερά η οποία καθορίζει το μέγεθος της ποινής, και γ(g) είναι η βαθμολογία του κενού συναρτήσει του μήκους του.

Έτσι στο προηγούμενο παράδειγμα, θα είχαμε βαθμολογία -2 και -3 για τα δύο κενά μήκους 2 και 3 αντίστοιχα.

Dot plots

Ο απλούστερος τρόπος για την απεικόνιση των ομοιοτήτων ανάμεσα σε δύο σειρές είναι με τη χρήση των **dot plots** (Philips, 1970). Τα dot plots είναι δισδιάστατοι πίνακες, στους οποίους η μία προς σύγκριση σειρά τοποθετείται στον οριζόντιο και η δεύτερη σειρά στον κάθετο άξονα. Εν συνεχεία, οι δύο σειρές συγκρίνονται και για κάθε όμοιο στοιχείο τους το αντίστοιχο κελί του πίνακα σημαδεύεται (βλ. *πίνακα 1.1*). Κατ' αυτόν τον τρόπο τα κοινά τμήματα στις δύο σειρές αναπαριστόνται σαν διαγώνιες γραμμές στον πίνακα.

Μπορεί να γίνει εύκολα κατανοητό, ότι σε περίπτωση απόλυτης ταύτισης των δύο σειρών, ο πίνακας θα διατρέχεται από την κεντρική διαγώνιο στο μέσο του. Οι τυχόν εισαγωγές και διαγραφές που υπάρχουν στις σειρές, στον πίνακα εμφανίζονται σαν διακοπές στη διαγώνιο. Επίσης, περιοχές των σειρών όπου παρατηρείται τοπική αντιστοίχηση, καθώς και επαναλαμβανόμενες σειρές, αναπαριστόνται με μικρότερες διαγωνίους στον πίνακα.

tuples

Εξαιτίας του περιορισμένου αλφαβήτου τόσο των αμινοξέων, όσο και των βάσεων, υπάρχει περίπτωση αντιστοιχήσεις να εμφανίζονται τυχαία. Για να μειωθεί το φαινόμενο αυτό (ο *θόρυβος*) μπορούμε να σημαδεύουμε όχι αντιστοιχίσεις μεμονομένων βάσεων, αλλά των λεγόμενων ***tuples***. Ως *k-tuple* χαρακτηρίζονται μία σειρά από *k residues* μέσα σε μία σειρά. Για παράδειγμα, ένα *2-tuple* αντιστοιχεί σε δύο συνεχόμενα residues.

	a	a	g	t	c	c	c	g	t	g
a	*	*								
g			*					*		*
g			*					*		*
t				*					*	
c					*	*	*			
c					*	*	*			
g			*					*		*
t				*					*	
t				*					*	
c					*	*	*			

	a	a	g	t	c	c	c	g	t	g
a		*								
g			*							
g			*					*		
t				*					*	
c					*	*				
c					*	*	*			
g			*					*		
t				*					*	
t				*						
c					*					

Πίνακας 1.1: (i) Dot plot *όπου αναπαρίσταται η αντιστοίχηση δύο σειρών. Τα κελιά του πίνακα τα οποία σχετίζονται με αντιστοιχισμένα στοιχεία των δύο σειρών έχουν σημαδευτεί.* ***(ii)*** *Μόνο τα κελιά τα οποία αντιστοιχούν σε tuples δύο και τεσσάρων βάσεων έχουν σημαδευτεί και έχει επισημανθεί το βέλτιστο μονοπάτι - αντιστοίχηση.*

Τα dot plots δεν παρέχουν αποτελέσματα μεγάλης ακριβείας, καθώς δεν προσφέρουν τη δυνατότητα ανεύρεσης της βέλτιστης αντιστοίχησης ανάμεσα σε δύο σειρές. Για να επιτευχθεί κάτι τέτοιο πρέπει να βρεθεί το βέλτιστο μονομάτι μέσα στο dot plot, και για το σκοπό αυτό χρησιμοποιούνται οι πίνακες αντικατάστασης, οι οποίοι αποτελούν μία παραλλαγή των dot plots.

Πίνακες αντικατάστασης (Substitution matrices)

Οι πίνακες αντικατάστασης ή πίνακες βαθμολογίας (**score matrices**) παρέχουν πληροφορίες σχετικά με αμινοξέα (residues) τα οποία χαρακτηρίζονται από παρόμοιες ιδιότητες, ή αμινοξέα τα οποία εναλλάσσονται τακτικά σε γνωστές οικογένειες πρωτεϊνών. Κάθε πίνακας δηλαδή, αναπαριστά κατά κάποιον τρόπο, μία ξεχωριστή εξελικτική θεωρία.

Δεδομένου αυτών των πληροφοριών τους, οι πίνακες αντικατάστασης χρησιμοποιούνται προκειμένου να 'κρατούν' τη βαθμολόγηση των αντικαταστάσεων που συναντώνται σε μία αντιστοίχηση, και ειδικά σε περιπτώσεις αντιστοίχησης σειρών πρωτεϊνών. Αυτό όμως δε σημαίνει ότι δεν μπορούν να χρησιμοποιηθούν και με σειρές DNA.

Οι τιμές των στοιχείων που αποθηκεύονται σε ένα πίνακα αντικατάστασης αναπαριστούν είτε την **ομοιότητα** – πόσο 'κοντινό' είναι δηλαδή ένα αμινοξύ με αυτό που αντικατέστησε στη σειρά – είτε

την **απόσταση** – ποιο είναι το κόστος από την αντικατάσταση ενός αμινοξέως με ένα άλλο. Η λογική πίσω και από τις δύο αυτές προσεγγίσεις είναι οι ίδια, επομένως οι πίνακες αντικατάστασης θα έχουν σχετικά σταθερή μορφή.

Η πιο συνήθης προσέγγιση για την εύρεση των τιμών ενός πίνακα αντικατάστασης είναι με τη χρήση πιθανοτήτων:

$$\mathbf{S_{ij} = \log (q_{ij} / p_i p_j)}$$

Όπου S_{ij} είναι η τιμή του (i, j) στοιχείου του πίνακα S και προκύπτει ως ο λογάριθμος του λόγου δύο πιθανοτήτων:

- $\mathbf{q_{ij}}$ είναι η πιθανότητα τα αμινοξέα i και j να έχουν προκύψει από έναν κοινό πρόγονο, επομένως να υπάρχει αντιστοιχία μεταξύ τους λόγω εξελικτικής συγγένειας

- $\mathbf{p_i}$ και $\mathbf{p_j}$ είναι οι πιθανότητες τυχαίας εμφάνισης των αμινοξέων i και j αντιστοίχως, επομένως το γινόμενο $\mathbf{p_i p_j}$ συμβολίζει την πιθανότητα να υπάρξει τυχαία αντιστοίχησή τους.

Οι πιο ευρέως χρησιμοποιούμενοι πίνακες αντικατάστασης είναι οι πίνακες **PAM** (Dayhoff, Schwarz και Orcutt - 1978) και οι πίνακες **BLOSUM** (Henikoff και Henikoff - 1991), οι οποίοι και οι δύο χρησιμοποιούνται σε αντιστοιχήσεις πρωτεϊνών.

Πίνακες PAM

Οι πίνακες PAM κατασκευάστηκαν λαμβάνοντας υπ' όψην 71 οικογένειες πρωτεϊνών, στις οποίες η ομοιότητα των πρωτεϊνών αγγίζει το 85% (δηλαδή, οι σειρές των πρωτεϊνών διαφέρουν το πολύ κατά το 15% των residues τους). Πραγματοποιώντας την αντιστοιχία των πρωτεϊνών αυτών, οι Dayhoff, Schwarz και Orcutt 'έχτισαν' ένα θεωρητικό φυλογενετικό δένδρο (phylogenetic tree – μία γραφική απεικόνιση των εξελικτικών σχέσεων μίας ομάδας οργανισμών) και προέβλεψαν τα residues τα οποία έχουν τη μεγαλύτερη πιθανότητα να εμφανιστούν στις προγονικές σειρές.

Η κατασκευή των πινάκων PAM βασίστηκε σε 1572 αλλαγές residues, και καταγράφει τη συχνότητα αντικατάστασης ενός residue X από ένα residue Y μέσα σε χρόνο Z, αγνοώντας την εξελικτική κατεύθυνση. Ο πρώτος πίνακας PAM ονομάστηκε 1PAM, καθώς απευθυνόταν σε σειρές όπου ο αριθμός των αποδεκτών μεταλλάξεων σε αυτές αποτελούσε το 1% του συνολικού μήκους τους. Προκειμένου να αυξηθεί η επιτρεπόμενη απόσταση, ο πίνακας PAM1 μπορεί να πολλαπλασιαστεί και να χρησιμοποιηθούν τα πολλαπλάσιά του. Η πιο διαδεδομένη έκδοση που χρησιμοποιείται είναι ο PAM250.

Χαρακτηριστικό των πινάκων PAM είναι ότι λειτουργούν καλά με σειρές οι οποίες έχουν κοντινή σχέση. Επίσης, βασίζονται σε δεδομένα όπου οι πιο συνήθεις αντικαταστάσεις είναι αλλαγές μίας μόνο βάσης σε κάποιο κωδικόνιο (codon). Μειονέκτημα των πινάκων αυτών αποτελεί το γεγονός ότι κλίνουν προς συντηρητικές μεταλλάξεις σε σειρές DNA, αντί για αντικαταστάσεις αμινοξέων, και οι οποίες μεταλλάξεις, όντας συντηρητικές, δεν επηρεάζουν σε μεγάλο βαθμό την λειτουργία και τη δομή της σειράς. Τέλος, η κάθε αντικατάσταση στη σειρά βασίζεται μόνο στο αμινοξύ το οποίο υπάρχει στη συγκεκριμένη θέση και στην πιθανότητα η οποία δίδεται από τον πίνακα. Αυτή η παραδοχή δεν αντικατοπτρίζει όμως σωστά τις διαδικασίες τις εξέλιξης.

Πίνακες BLOSUM

Η άλλη κατηγορία πινάκων αντικατάστασης που χρησιμοποιούνται συχνά είναι οι πίνακες BLOSUM. Για την κατασκευή τους χρησιμοποιήθηκαν σύνολα περιοχών χωρίς κενά. Οι περιοχές αυτές ανήκουν σε οικογένειες πρωτεϊνών που περιέχονται στη βάση BLOCKS. Η βάση δεδομένων BLOCKS περιλαμβάνει ομάδες (clusters) από σύντομες σειρές πρωτεϊνών οι οποίες παρουσιάζουν μεγάλη ομοιότητα. Οι ομάδες αυτές προκύπτουν από τη βάση SWISS-PROT και άλλες βάσεις,

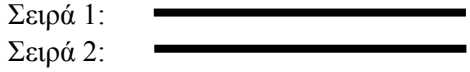
εφαρμόζοντας σε αυτές τον αλγόριθμο MOTIF. Η ομαδοποίηση των σειρών γίνεται θέτοντας κάποιο συγκεκριμένο ποσοστό ομοιότητας για τις σειρές ως όριο και τοποθετώντας στην ίδια ομάδα σειρές που υπερβαίνουν το συγκεκριμένο ποσοστό. Επίσης, υπολογίζεται η συχνότητα με την οποία δύο residues τα οποία έχουν αντιστοιχιστεί σε μία ομάδα να τύχει να αντιστοιχιστούν και σε μία άλλη.

Χρησιμοποιώντας όλες τις άνωθεν παρατηρήσεις, το αποτέλεσμα που προκύπτει είναι ο λόγος των κατεγραμμένων αντικαταστάσεων μεταξύ δύο οποιονδήποτε residues προς όλες τις αντικαταστάσεις που έχουν καταγραφεί. Όπως και με τους πίνακες PAM, υπάρχουν πολλές εκδόσεις των πινάκων BLOSUM, με τη διαφορά ότι η αρίθμησή τους είναι αντίστροφη από αυτή των πινάκων PAM. Έτσι, π.χ. ο πίνακας BLOSUM50 περιλαμβάνει ομάδες σειρών με τουλάχιστον 50% ομοιότητα, ενώ ο BLOSUM62 περιλαμβάνει ομάδες σειρών με τουλάχιστον 62% ομοιότητα.

Αλγόριθμοι αντιστοίχησης (Alignment algorithms)

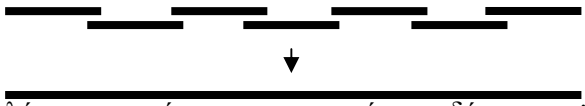
Έχοντας καταλήξει σχετικά με το σύστημα βαθμολόγησης που θα χρησιμοποιηθεί, το επόμενο βήμα είναι η εύρεση του βέλτιστου τρόπου αντιστοίχησης των δύο σειρών. Οι τρόποι με τους οποίους μπορεί να πραγματοποιηθεί η αντιστοίχηση είναι οι ακόλουθοι:

- **Καθολική αντιστοίχηση** δύο σειρών (**global alignment**). Η αντιστοίχηση αυτή χρησιμοποιείται σε περιπτώσεις όπου οι σειρές έχουν ακριβώς ή σχεδόν το ίδιο μήκος.



- **Τοπική αντιστοίχηση** δύο σειρών (**local alignment**). Χρησιμοποιείται για την εύρεση κοινών υποσειρών μέσα στις σειρές.

- **Αντιστοίχηση με ελεύθερα άκρα (Ends free alignment)**. Για την εύρεση ενώσεων (joins) / επικαλύψεων (overlaps).



Δυναμικός Προγραμματισμός (Dynamic Programming)

Ένα από τα δυσκολότερα σημεία στην αντιστοίχηση δύο πρωτεϊνών είναι η αναγνώριση των εισαγωγών ή των διαγραφών σε κάθε σειρά. Η αντιστοίχηση δύο σειρών βελτιώνεται σε μεγάλο βαθμό και γίνεται πιο ακριβής με την εισαγωγή των καταλλήλων κενών (τα οποία θα αντιστοιχούν στις εισαγωγές και τις διαγραφές). Η σωστή τοποθέτηση των κενών όμως είναι μία δύσκολη διαδικασία η οποία δεν μπορεί να γίνει τυχαία, ειδικά σε περιπτώσεις μεγάλων σειρών. Επιπλέον, σε περιπτώσεις αντιστοίχησης πολύ μακρινών ομολόγων σειρών, τα κενά μπορεί να είναι πολύ μεγάλου μήκους. Υπάρχουν συγκεκριμένες, ακριβείς μέθοδοι οι οποίες χρησιμοποιούν μεθόδους βελτιστοποίησης, για την αναγνώριση και τοποθέτηση των κενών στα σωστά σημεία, έτσι ώστε να επιτυγχάνεται η σωστότερη δυνατή αντιστοίχηση.

Ένα άλλο ευαίσθητο σημείο, είναι ότι για δύο σειρές ιδίου μήκους *n*, ακόμα και για μικρές τιμές του, υπάρχουν πολλοί πιθανοί συνδυασμοί αντιστοιχήσεων τους, επομένως είναι δύσκολο να υπολογιστούν όλοι.

Δεδομένης της χρήσης μίας αθροιστικής μεθόδου βαθμολόγησης, όπως αυτή ήδη περιγράφηκε, στην αντιστοίχηση δύο σειρών, ο αλγόριθμος που χρησιμοποιείται σε τέτοιες περιπτώσεις για την εύρεση της καλύτερης αντιστοίχησης βασίζεται στο **δυναμικό προγραμματισμό (dynamic programming)**. Επίσης, ο δυναμικός προγραμματισμός εφαρμόζει μία μέθοδο βελτιστοποίησης, έτσι ώστε να διαχειρίζεται επιτυχώς τις εισαγωγές και διαγραφές σε μία σειρά.

Η απλούστερη μορφή αλγορίθμου δυναμικού προγραμματισμού για την αντιστοίχηση σειρών είναι η **αντιστοίχηση σειρών ανά ζεύγη** (**pair-wise sequence alignment**). Οι αλγόριθμοι δυναμικού προγραμματισμού εγκυόνται να βρουν τη βέλτιστη αντιστοίχηση με βάση τη βαθμολογία που αποδίδεται σε κάθε αντιστοίχηση. Εκτός από τους αλγορίθμους δυναμικού προγραμματισμού, μπορούν να εφαρμοστούν και ευρεστικές μέθοδοι (heuristic methods) για να εκτελέσουν την ίδια αναζήτηση. Οι μέθοδοι αυτές είναι ταχύτατες, με το μειονέκτημα όμως ότι θέτουν επιπλέον παραδοχές, με αποτέλεσμα σε ορισμένες περιπτώσεις να υπάρχει κίνδυνος να χάσουν κάποια καλή αντιστοίχηση. Θα αναφερθούμε λίγο πιο αναλυτικά στις ευρεστικές μεθόδους στη συνέχεια.

Όπως προαναφέρθηκε, ο τρόπος βαθμολόγησης που θα χρησιμοποιήσουμε για τις αντιστοιχίσεις είναι αθροιστικός. Επομένως, ο αλγόριθμος δυναμικού προγραμματισμού που θα εφαρμοστεί, θα στοχεύει στο να βρει την αντιστοίχηση με τη μεγαλύτερη δυνατή βαθμολογία. Υπάρχουν περιπτώσεις κατά τις οποίες στη σύγκριση βιολογικών σειρών χρησιμοποιείται διαφορετικός τρόπος βαθμολόγησης και η βαθμολογία που αποδίδεται σε κάθε αντιστοίχηση μεταφράζεται ως ποινή ή κόστος. Σε τέτοιες περιπτώσεις ο αλγόριθμος δυναμικού προγραμματισμού θα λειτουργήσει με τον ίδιο τρόπο. Η μόνη διαφορά είναι ότι το ζητούμενο πλέον θα είναι να ελαχιστοποιηθεί η βαθμολογία – ποινή – κάθε αντιστοίχησης, αντί να μεγιστοποιηθεί.

Αλγόριθμος Needleman - Wunsch

Έχοντας δύο σειρές ιδίου ή σχεδόν ιδίου μήκους, το ζητούμενο είναι να βρούμε τη βέλτιστη καθολική αντιστοίχηση (global alignment), επιτρέποντας και τα κενά. Ο αλγόριθμος Needleman – Wunsch είναι ο αλγόριθμος δυναμικού προγραμματισμού ο οποίος χρησιμοποιείται σε τέτοιες περιπτώσεις. Η βασική ιδέα του αλγορίθμου είναι να χτιστεί η βέλτιστη αντιστοίχηση χρησιμοποιώντας προηγούμενες λύσεις από βέλτιστες αντιστοιχίσεις μικρότερων υποσειρών.

Τα κύρια σημεία του αλγορίθμου είναι ότι βρίσκει τη βέλτιστη αντιστοίχηση δύο σειρών, βασιζόμενος στη βέλτιστη βαθμολογία, συμπεριλαμβάνοντας ΟΛΕΣ τις βάσεις και από τις δύο σειρές. Επίσης, τα κενά προστίθονται στο εσωτερικό, ή στα άκρα κάθε σειράς με αποτέλεσμα το μήκος των δύο σειρών (βάσεις + κενά) να είναι ακριβώς το ίδιο. Τέλος, κάθε βάση ή κενό στην κάθε σειρά αντιστοιχίζεται με μία βάση ή κενό στην άλλη σειρά.

Ας υποθέσουμε τώρα πως έχουμε δύο σειρές *S* και *T*.

- Η σειρά *S* αποτελείται από *n* βάσεις, επομένως έχει μήκος *n*, και η σειρά *T* αποτελείται από *m* βάσεις (έχει λοιπόν μήκος *m*).

- Θέτουμε μία αυθαίρετη ποινή για τα κενά της τάξης του -1 για κάθε indel.

- Συμβολίζουμε την αντιστοίχηση μεταξύ της βάσης *i* στη σειρά *S* με ένα κενό στη σειρά *T* ως: (S_i, -).

- Η βαθμολογία στην περίπτωση αυτή συμβολίζεται ως: $\sigma(S_i, -) = -1$.

- Συμβολίζουμε την αντιστοίχηση μεταξύ της βάσης *i* στη σειρά *S* με τη βάση *j* στη σειρά *T* ως: (S_i, T_j).

- Η βαθμολογία στην περίπτωση αυτή συμβολίζεται ως: $\sigma(S_i, T_j) = -1$.

	A	**C**	**G**	**T**
A	2	-1	-1	-1
C	-1	2	-1	-1
G	-1	-1	2	-1
T	-1	-1	-1	2

Πίνακας 1.1: *Ένα απλό παράδειγμα πίνακα αντικατάστασης μεταξύ δύο πανομοιότυπων σειρών, αποδίδοντας 2 βαθμούς για κάθε ακριβή αντιστοίχηση βάσης και -1 βαθμό σε κάθε κενό ή mismatch.*

- Για πίνακα αντικατάστασης, φτιάχνουμε ένα πίνακα διαστάσεων n+1 επί m+1.

- Η γραμμή 0 και η στήλη 0 του πίνακα αναπαριστούν το κόστος που θα είχαμε προσθέτοντας διαδοχικά κενά και στις δύο σειρές κατά την έναρξη της αντικατάστασης.

- Η βαθμολογία για κάθε κελί του πίνακα υπολογίζεται αναδρομικά, με βάση τις βαθμολογίες των γύρω και προηγούμενων από αυτό κελιών, βρίσκοντας τη βέλτιστη επιλογή ανάμεσα σε αυτά η οποία αναπαριστά είτε κενό, είτε επιτυχία / αποτυχία αντιστοίχησης.

Είναι πιο εύκολο να χρησιμοποιήσουμε ένα παράδειγμα του αλγορίθμου, παρά να επιχειρήσουμε να τον περιγράψουμε λεπτομερώς.

Ας ξεκινήσουμε λοιπόν με τις ακόλουθες δύο σειρές :

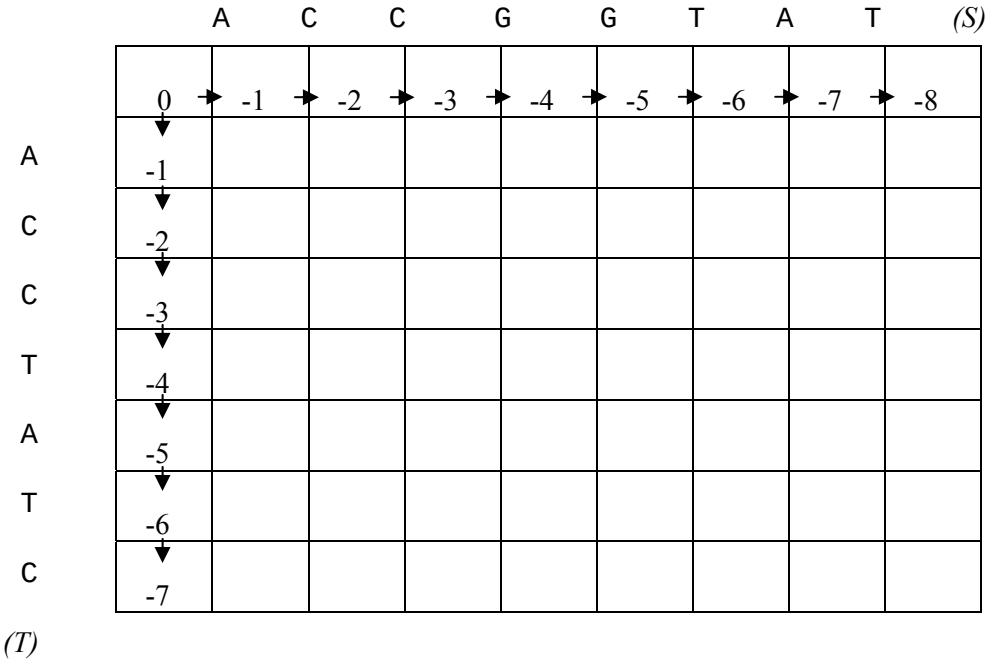
$S =$ ACCGGTAT

$T =$ ACCTATC

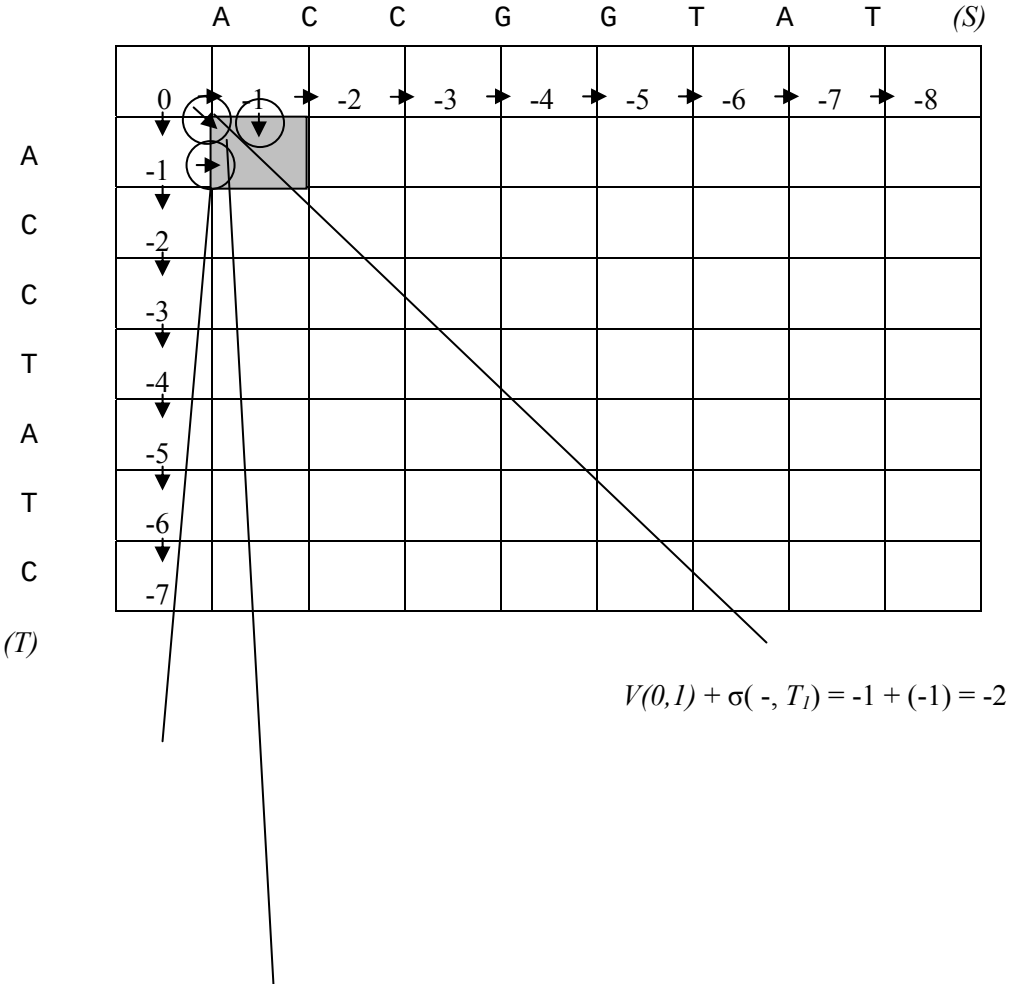
Το μήκος της σειράς *S* είναι $n = 8$ και της σειράς *T* είναι $m = 7$. Τα δύο μήκη είναι σχεδόν τα ίδια, επομένως μπορούμε να χρησιμοποιήσουμε καθολική αντιστοίχηση.

Φτιάχνουμε τον πίνακα *V* διαστάσεων (n+1=)9 επί (m+1=)8. Ξεκινάμε, ορίζοντας την τιμή του στοιχείου $V(0,0) = 0$ και γεμίζουμε τον υπόλοιπο πίνακα κινούμενοι γραμμή – γραμμή από πάνω αριστερά προς κάτω δεξιά. Γνωρίζοντας τις τιμές των $V(i-1, j-1)$, $V(i-1, j)$ και $V(i, j-1)$ μπορεί να υπολογιστεί η τιμή του $V(i, j)$, η οποία δίδεται ως εξής:

$$V(i,j) = \max \begin{cases} V(i-1, j-1) + \sigma(S_i, T_j) \\ V(i-1, j) + \sigma(S_i, -) \\ V(i, j-1) + \sigma(-, T_j) \end{cases}$$



Πίνακας 1.2: *Πίνακας αντικατάστασης διαστάσεων 9x8 για τις σειρές S =* ACCGGTAT *και T =* ACCTATC*. Όπως αναφέρθηκε, η πρώτη γραμμή και η πρώτη στήλη συμπληρώνονται σαν να προσθέταμε διαδοχικά κενά και στις δύο σειρές.*



$V(1,0) + \sigma(S_1, -) = -1 + (-1) = -2$

Κινούμενοι καθέτως, θεωρούμε <u>πάντα</u> (ακόμα και αν υπάρχει ταύτιση των βάσεων) ότι εισάγουμε κενό στη σειρά *S* η οποία αναπαρίσταται οριζόντια στον πίνακα. Επομένως, η βαθμολογία που προστίθεται σε αυτή του στοιχείου *V(i, j-1)* είναι <u>πάντα</u> η ποινή που αντιστοιχεί σε κενό και που στο συγκεκριμένο παράδειγμα είναι -1. Το ίδιο ισχύει πάντα και όταν κινούμαστε οριζοντίως, με τη μόνη διαφορά ότι το κενό εισάγεται στην κάθετη σειρά *T*.

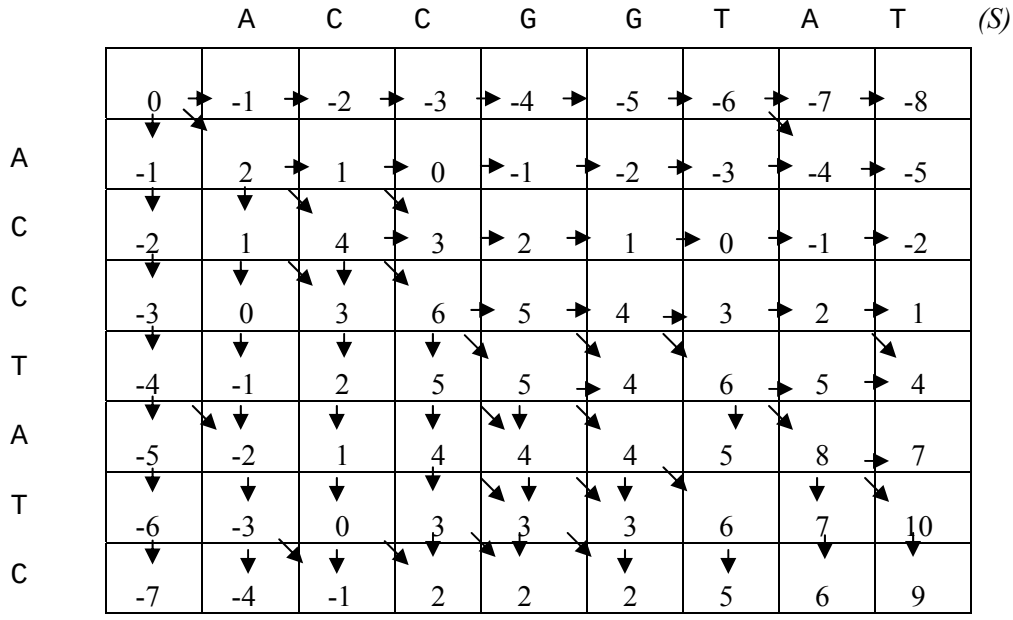
$V(1,1) + \sigma(S_1, T_1) = 0 + 2 = 2$

Σε περίπτωση που υπάρχει ακριβής ταύτιση των αντιστοιχούμενων βάσεων μεταξύ των δύο σειρών, κινούμενοι <u>μόνο</u> διαγωνίως, προσθέτουμε τον προκαθορισμένο βαθμό (στην προκειμένη $\sigma(S_i, T_i) = 2$) στην τιμή του στοιχείου *V(i-1, j-1)*.

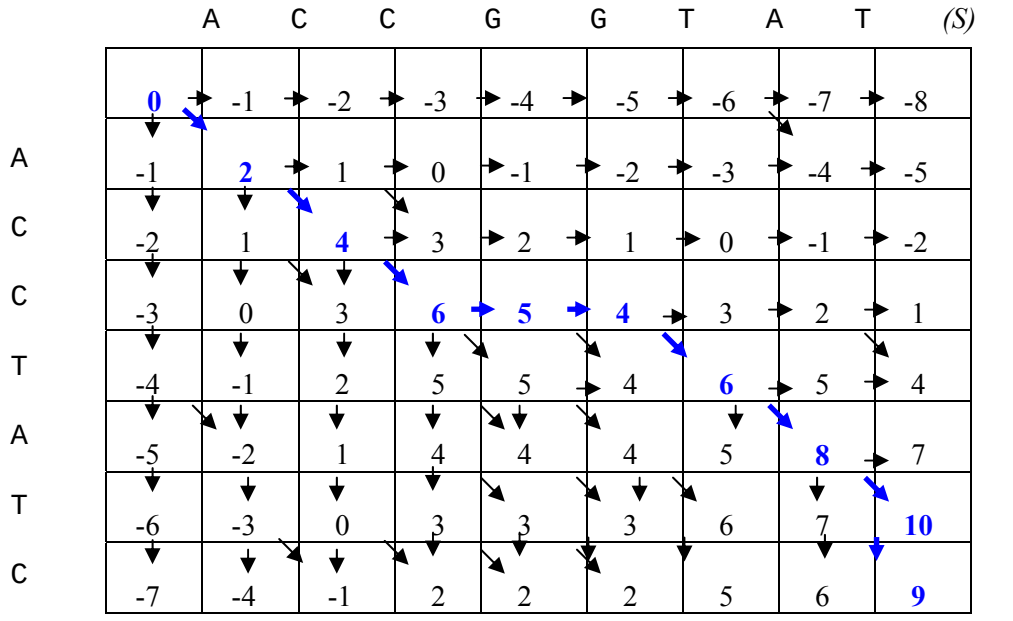
Πίνακας 1.3: *Στη συνέχεια, για κάθε νέο στοιχείο του πίνακα βρίσκουμε το 'βέλτιστο' μονοπάτι. Η βέλτιστη επιλογή είναι αυτή που αποδίδει τη μεγαλύτερη βαθμολογία και στην προκειμένη περίπτωση είναι 2.*

Σε περίπτωση μη ταύτισης μπορούμε να κινηθούμε είτε διαγωνίως, είτε οριζοντίως ή καθέτως. Κινούμενοι διαγωνίως, προσθέτουμε την (αρνητική) ποινή που έχει οριστεί για τη μη-ταύτιση, ενώ κινούμενοι οριζοντίως ή καθέτως προσθέτουμε την (αρνητική) ποινή που έχει οριστεί για το κάθετο ($\sigma(S_i, -)$), ή το οριζόντιο ($\sigma(-, T_i)$) κενό. Σε περίπτωση ταύτισης, μπορούμε πάλι να κινηθούμε είτε διαγωνίως, είτε οριζοντίως ή καθέτως, μόνο που τώρα κινούμενοι διαγωνίως προσθέτουμε τη (θετική) βαθμολογία $\sigma(S_i, T_i)$, ενώ κινούμενοι οριζοντίως ή καθέτως προσθέτουμε και πάλι την ποινή, καθώς αυτή η κίνηση συμβολίζει <u>πάντα</u> κενό στην κάθετη ή στην οριζόντια σειρά αντιστοίχως.

Κάθε φορά που αποδίδεται μία τιμή σε ένα στοιχείο του πίνακα, αποθηκεύεται και ο δείκτης ο οποίος δείχνει από ποιο προηγούμενο στοιχείο οδηγηθήκαμε στο παρόν. Κατ' αυτόν τον τρόπο μπορούμε να κινηθούμε αναδρομικά και να κατασκευάσουμε τις δύο αντιστοιχισμένες σειρές, πλήρεις με τα κενά που τους έχουν αποδοθεί. Εάν μία τιμή ενός στοιχείου έχει προκύψει από τα δύο ή και τα τρία προηγούμενα στοιχεία του πίνακα, οι δείκτες απ' όλα τα στοιχεία κρατούνται και έτσι προκύπτουν αντίστοιχα δύο ή τρία διαφορετικά μονοπάτια. Κάθε μονοπάτι απεικονίζει και μία διαφορετική αντιστοίχηση και η επιλογή του βέλτιστου μονοπατιού / αντιστοίχησης, γίνεται πλέον με γνώμονα τη μέγιστη βαθμολόγηση που αντιστοιχεί σε κάποιο από αυτά, είτε σε περίπτωση που τα μονοπάτια είναι ισότιμα, η επιλογή γίνεται αυθαίρετα .



Πίνακας 1.4: *Ο συμπληρωμένος πίνακας αντικατάστασης για τις σειρές S = ACCGGTAT και T = ACCTATC. Οι δείκτες δείχουν από ποιο προηγούμενο στοιχείο του πίνακα έχει προκύψει η τιμή του παρόντος.*



Πίνακας 1.5: *Ξεκινώντας από το ακρότατο σημείο κάτω και δεξιά, κινούμενοι αναδρομικά, και οδηγούμενοι από τους δείκτες, 'χτίζουμε' τις δύο σειρές μέχρι να φτάσουμε στο πάνω αριστερά άκρο. Οι δύο σειρές σε αντιστοιχία είναι:*

```
ACCGGTAT- (S)
|||   |||
ACC--TATC (T)
```

Στην αντιστοίχηση υπάρχουν: 6 ταυτίσεις = 6x2 = 12 βαθμοί, 3 κενά = 3x(-1) = -3 βαθμοί. Επομένως η συνολική βαθμολογία είναι: 12+(-3) = 9 βαθμοί.

Ο λόγος που η συγκεκριμένη μέθοδος αντιστοίχησης είναι επιτυχής, είναι ότι η βαθμολόγηση απαρτίζεται από το άθροισμα ανεξάρτητων μεταξύ τους τμημάτων. Έτσι, η βέλτιστη βαθμολογία για κάθε δεδομένη στιγμή της αντιστοίχησης είναι η βέλτιστη βαθμολογία του αμέσως προηγούμενου βήματος του αλγορίθμου, συν ο βαθμός που αποδίδεται στο παρόν βήμα.

Αξιολόγηση του αλγορίθμου Needleman - Wunsch

Για την αξιολόγηση του αλγορίθμου Needleman – Wunsch, μπορούμε να καθορίσουμε την αποδοτικότητα του ως προς το χρόνο επεξεργασίας και τη μνήμη αποθήκευσης. Λαμβάνοντας υπ' όψην ότι αποθηκεύουμε *(n+1)x(m+1)* αριθμούς (όλα τα στοιχεία του πίνακα αντικατάστασης), και κάθε ένας από αυτούς τους αριθμούς απαιτεί ένα συγκεκριμένο αριθμό υπολογισμών (τρία αθροίσματα και ένα μέγιστο). Η πολυπλοκότητα του αλγορίθμου θα είναι επομένως *O(nm)* ως προς τον απαιτούμενο χρόνο και *O(nm)* ως προς την απαιτούμενη μνήμη. Όσο μεγαλύτερες είναι οι τιμές των *n* και *m*, τόσο μεγαλώνει και η πολυπλοκότητα του αλγορίθμου και μειώνεται η αποδοτικότητά του. Καθώς οι βιολογικές σειρές είναι συνήθως μεγάλου μήκους, ο αλγόριθμος Needleman – Wunsch αποδεικνύεται εφικτός και αποδίδει καλά αποτελέσματα, αλλά είναι σχετικά αργός.

Αλγόριθμος Smith – Waterman

Μέχρι στιγμής ασχοληθήκαμε με την καθολική αντιστοίχηση δύο σειρών. Τι συμβαίνει όμως σε περιπτώσεις που κρίνεται άσκοπη μία κακή αντιστοίχηση μεταξύ δύο ολόκληρων σειρών και που κρίνεται προτιμότερο να βρούμε τη βέλτιστη αντιστοίχηση ανάμεσα σε *υποσειρές* δύο σειρών; Κάτι τέτοιο κρίνεται σκόπιμο, όταν π.χ. υπάρχει περίπτωση δύο πρωτεΐνες να έχουν μία κοινή περιοχή (domain). Σε αυτήν την περίπτωση χρησιμοποιούμε την **τοπική αντιστοίχηση (local alignment)**.

Σε γενικές γραμμές, η τοπική αντιστοίχηση θεωρείται πολύ πιο ευαίσθητη μέθοδος αντιστοίχησης δύο σειρών και είναι χρήσιμη σε περιπτώσεις που δύο σειρές, αν και ομόλογες, έχουν διαφοροποιηθεί πολύ μέσα στην εξελικτική τους πορεία. Τότε μόνο ορισμένες μόνο περιοχές των σειρών θα μπορούν να ταυτιστούν, καθώς οι υπόλοιπες θα έχουν διαφοροποιηθεί τόσο πολύ εξαιτίας του προσαρτώμενου σε αυτές θορύβου.

Η τοπική αντιστοίχηση παρουσιάζει πολλά κοινά με την ολική: χρησιμοποιεί τον ίδιο τύπο πινάκων αντικατάστασης και ποινών για τα κενά, ενώ κάνει χρήση και αλγορίθμων δυναμικού προγραμματισμού, με ορισμένες μόνο παραλλαγές.

Ο αλγόριθμος δυναμικού προγραμματισμού που χρησιμοποιείται για την τοπική αντιστοίχηση είναι ο **Smith – Waterman**. Ο πίνακας αντικατάστασης για τον αλγόριθμο Smith – Waterman κατασκευάζεται με τον ίδιο τρόπο όπως και για τον αλγόριθμο Needleman – Wunsch, με μόνη διαφορά ότι στον πίνακα αυτό η κατώτερη τιμή που μπορεί να αποθηκευτεί για κάποιο στοιχείο του είναι 0.

Έτσι, για έναν πίνακα αντικατάστασης V και γνωρίζοντας τις τιμές των $V(i-1, j-1)$, $V(i-1, j)$ και $V(i, j-1)$ οι τιμές των στοιχείων του προκύπτουν ως εξής:

$$V(i, 0) = 0,\ V(0, j) = 0, \text{ για κάθε } i, j$$

$$V(i, j) = \max \begin{cases} 0 \\ V(i-1, j-1) + \sigma(S_i, T_j) \\ V(i-1, j) + \sigma(S_i, -) \\ V(i, j-1) + \sigma(-, T_j) \end{cases}$$

Η επιλογή της απόδοσης σε ένα στοιχείο του πίνακα την τιμή 0, αντί μίας αρνητικής τιμής, σηματοδοτεί στην έναρξη μίας νέας αντιστοιχίας. Η λογική είναι ότι σε περίπτωση που προκύψει αρνητικός βαθμός σε μία αντιστοίχηση είναι προτιμότερο να ξεκινήσει μία νέα αντιστοιχία υποσειρών αντί να συνεχίσει η προηγούμενη.

Με τη συμπλήρωση του πίνακα αντικατάστασης, βρίσκουμε το στοιχείο (σε οποιαδήποτε θέση του πίνακα) το οποίο έχει την υψηλότερη τιμή ανάμεσα στα στοιχεία του πίνακα. Με αφετηρία το στοιχείο αυτό και ακολουθώντας τους δείκτες, κινούμαστε αναδρομικά, μέχρις ότου συναντήσουμε στοιχείο με την τιμή 0 και το οποίο μπορεί να βρίσκεται σε οποιαδήποτε θέση του πίνακα. Η ολική βαθμολογία που αποδίδεται στην αντιστοίχηση υποσειρών είναι η τιμή του στοιχείου απ' όπου ξεκινήσαμε την αναδρομή (η υψηλότερη τιμή στοιχείου στον πίνακα).

Σε ορισμένες περιπτώσεις ενδέχεται το αποτέλεσμα της τοπικής αντιστοίχησης μεταξύ δύο σειρών να είναι υποσύνολο της ολικής αντιστοίχησης των ιδίων σειρών, όμως κάτι τέτοιο δεν πρέπει να θεωρείται πάντα δεδομένο.

		A	C	C	G	G	T	A	T	*(S)*
	0	0	0	0	0	0	0	0	0	
T	0	0	0	0	0	0	2	1	2	
T	0	0	0	0	**0**	0	2	1	3	
G	0	0	0	0	2	**2**	1	1	2	
T	0	0	0	0	1	1	**4**	3	3	
A	0	2	1	0	0	0	3	**6**	5	
T	0	1	1	0	0	0	2	5	**8**	
C	0	0	3	3	2	1	1	4	7	
(T)										

Πίνακας 1.6: *Ο συμπληρωμένος πίνακας αντικατάστασης για την τοπική αντιστοίχηση των σειρών S = ACCGGTAT και T = TTGTATC. Ξεκινώντας από το στοιχείο του πίνακα με τη μεγαλύτερη τιμή (στην*

παρούσα περίπτωση είναι V(9,7) = 8) και κινούμενοι αναδρομικά ακολουθώντας τους δείκτες μέχρις ότου συναντήσουμε το V(5,3) = 0, βρίσκουμε τη βέλτιστη τοπική αντιστοίχηση των σειρών. Οι υποσειρές που προκύπτουν με βαθμολογία 8, είναι οι εξής:

```
GTAT  (S)
||||
GTAT  (T)
```

Όπως και με τον αλγόριθμο Needleman – Wunsch, έτσι και στον αλγόριθμό Smith – Waterman η πολυπλοκότητα είναι *O(nm)* ως προς τον απαιτούμενο χρόνο και *O(nm)* ως προς την απαιτούμενη μνήμη.

Ends – free αντιστοίχηση

Σε περιπτώσεις που η μία σειρά από τις δύο που θέλουμε να αντιστοιχίσουμε περιέχει την άλλη, ή ορισμένες περιοχές των δύο σειρών τυγχάνει να επικαλύπτονται, χρειάζεται διαφορετική προσέγγιση για την αντιστοίχηση των εν λόγω σειρών. Τέτοιες περιπτώσεις απαντώνται όταν συγκρίνουμε τμήματα σειρών DNA μεταξύ τους ή με μεγαλύτερες σειρές χρωμοσωμάτων.

Η προσέγγιση που χρησιμοποιείται στις παραπάνω περιπτώσεις αποτελεί μία ακόμα παραλλαγή της μεθόδου ολικής αντιστοίχησης σειρών. Ο πίνακας αντικαταστάσεων συμπληρώνεται με τον ίδιο αναδρομικό μοντέλο που εφαρμόστηκε και στην ολική / τοπική αντιστοίχηση. Για το βέλτιστο μονοπάτι εντοπίζουμε το ακραίο στοιχείο του πίνακα (είτε στην ακραία γραμμή του πίνακα είτε στην ακραία στήλη του) με την καλύτερη τιμή και χρησιμοποιώντας αυτό ως αφετηρία κινούμαστε αναδρομικά και ‘χτίζουμε’ την αντιστοιχία.

Για πίνακα αντικατάστασης *V* και γνωρίζοντας τις τιμές των *V(i-1, j-1)*, *V(i-1, j)* και *V(i, j-1)* οι τιμές των στοιχείων του προκύπτουν ως εξής:

$$V(i, 0) = 0,\ V(0, j) = 0, \text{ για κάθε } i, j$$

$$V(i, j) = \max \begin{cases} V(i-1, j-1) + \sigma(S_i, T_j) \\ V(i-1, j) + \sigma(S_i, -) \\ V(i, j-1) + \sigma(-, T_j) \end{cases}$$

		G	T	T	A	C	T	G	T	*(S)*
	0	0 →	0 →	0 →	0	0	0	0	0	
C	0	-1	-1	-1	-1	2 ↘	1	0	-1	
T	0	-1	1	1	0	1 ↘	4	3	2	
G	0	2	1	0	0	0	3 ↘	6	5	
T	0	1	4	4	3	2	2	5 ↘	8 ↓	
A	0	0	3	3	6	5	4	4	7 ↓	
T	0	-1	2	5	5	5	7	6	6 ↓	
C	0	-1	1	4	4	7	6	6	5	
(T)										

Πίνακας 1.7: *Ο συμπληρωμένος πίνακας αντικατάστασης για την ends-free αντιστοίχηση των σειρών S = GTTACTGT και T = CTGTATC. Ξεκινώντας από το ακραίο στοιχείο του πίνακα με τη μεγαλύτερη τιμή (στην παρούσα περίπτωση είναι το V(9,5) = 8) και κινούμενοι αναδρομικά ακολουθώντας τους δείκτες μέχρις ότου συναντήσουμε το επίσης ακραίο V(1,5) = 0, βρίσκουμε τη βέλτιστη ends-free αντιστοίχηση των σειρών. Η αντιστοίχηση επικάλυψης των σειρών που προκύπτει είναι:*

```
GTTACTGT---   (S)
    ||||
----CTGTATC   (T)
```

Affine Gap Penalty

Οι αλγόριθμοι δυναμικού προγραμματισμού που περιγράφηκαν μπορούν να δώσουν πολύ ικανοποιητικά αποτελέσματα για καθολικές, τοπικές και ends-free αντιστοιχίσεις. Στα παραδείγματα μας όμως χρησιμοποιήσαμε την απλή, γραμμική ποινή για τα τυχόν κενά στις σειρές, προσέγγιση η οποία γενικά δεν ενδείκνυται. Είναι προτιμότερο να χρησιμοποιηθεί κάποια άλλη, πιο διαδεδομένη μέθοδος, όπως για παράδειγμα η λεγόμενη **affine gap** ποινή. Σύμφωνα με αυτή τη μέθοδο, το κόστος για το 'άνοιγμα' ενός νέου κενού δεν υπολογίζεται το ίδιο όσο το κόστος της επέκτασης ενός ήδη υπάρχοντος κενού. Δε θα προχωρήσουμε όμως σε περαιτέρω ανάλυση της μεθόδου επιλογής των ποινών για τα κενά.

Ευρεστικές μέθοδοι (Heuristic methods)

Οι μέθοδοι δυναμικού προγραμματισμού δίνουν τα ακριβέστερα αποτελέσματα όσον αφορά την αντιστοίχηση και την ακριβή βαθμολογία σειρών, εντούτοις σε πραγματικές συνθήκες παρουσιάζουν ορισμένα μειονεκτήματα. Το πρόβλημα που προκύπτει οφείλεται στο μεγάλο μέγεθος των σειρών, το οποίο είναι ανάλογο με την πολυπλοκότητα και επομένως αυτομάτως αυξάνει τις απαιτήσεις σε πόρους χρόνου και μνήμης. Για την αντιμετώπιση του προβλήματος έχουν δοκιμαστεί διάφορες προσεγγίσεις, όπως για παράδειγμα η χρήση παραλλήλων υπολογιστών στους οποίους κατανέμονται οι διάφορες διαδικασίες και έτσι δεν εκτελούνται όλες από ένα μηχάνημα. Η συγκεκριμένη

προσέγγιση αν και λύνει ως ένα βαθμό τα προβλήματα του δυναμικού προγραμματισμού, αποδεικνύεται πολυδάπανη.

Μία άλλη προσέγγιση αποτελεί και η χρήση **ευρεστικών (heuristic) μεθόδων**, οι οποίες είναι γρηγορότερες και φθηνότερες από τους 'ακριβείς' αλγορίθμους δυναμικού προγραμματισμού, καθώς βασίζονται στο λογισμικό. Στο σημείο αυτό να σημειώσουμε ότι ευρεστικές ονομάζονται οι μέθοδοι επίλυσης προβλημάτων οι οποίες βασίζονται στην εμπειρία και κινούνται προς μία λύση οδηγούμενες από δοκιμές και τυχόν λάθη τους. Το κύριο μειονέκτημά τους είναι ότι δεν μπορούν να εγγυηθούν ότι θα καταλήξουν στην καλύτερη δυνατή αντιστοίχηση, καθώς τα αποτελέσματα τους είναι προσεγγιστικά. Θα δώσουν όμως μία καλή εικόνα της ομοιότητας των δύο προς σύγκριση σειρών, στηριζόμενες στην αναγνώριση ομοιοτήτων ανάμεσα σε τμήματά τους, αντί ολόκληρων των σειρών. Οι δύο πλέον διαδεδομένες ευρεστικές μέθοδοι είναι το FASTA και το BLAST, τα οποία θα περιγραφούν αναλυτικά στη συνέχεια.

FASTA

Σε μια βάση δεδομένων μπορούν να βρεθούν σειρές που είναι παρόμοιες με μία ζητούμενη σειρά. Αυτό επιτυγχάνεται, συγκρίνοντας τη ζητούμενη αλληλουχία με κάθε αλληλουχία που είναι αποθηκευμένη στη βάση δεδομένων και ανακτώντας τις σειρές με τη μεγαλύτερη βαθμολογία (δηλ. τις πιο όμοιες). Η μέθοδος που εφαρμόζεται για την αντιστοίχηση των αλληλουχιών είναι ο δυναμικός προγραμματισμός .

Το FASTA είναι ένα πρόγραμμα που εφαρμόζει ένα αλγόριθμο για εύρεση ομοιότητας, αυτός αλγόριθμος επικεντρώνεται στο να βρει περιορισμένου μήκους ομοιότητες (αποκαλούμενες λέξεις) μεταξύ δύο αλληλουχιών, και κατόπιν συνδυάζοντας αυτές τις ομοιότητες για να βρεθεί η συνολική ομοιότητα.

Τα αρχείο με τα αποτελέσματα από μια αναζήτηση σειράς πρωτεΐνης με το πρόγραμμα FASTA παρουσιάζεται παρακάτω.

Η ονομασία και η έκδοση του προγράμματος δίνονται στη κορυφή του αρχείου, με την κατάλληλη αναφορά για να χρησιμοποιηθεί σε οποιαδήποτε δημοσιευμένα αποτελέσματα που προέρχονται από αυτό το πρόγραμμα.

Η αναζητούμενη αλληλουχία δηλώνεται (π.χ. human alpha-1A adrenergic receptor), μαζί με την ονομασία της βάσης δεδομένων (SWISS_PROT) που γίνεται η αναζήτηση.

Μετά δίνονται πληροφορίες για τις παραμέτρους που χρησιμοποιήθηκαν στον αλγόριθμο και το χρόνο που χρειάστηκε το πρόγραμμα για να τρέξει (24.750’’).

Στη συνέχεια, ακολουθούν τα αποτελέσματα από την αναζήτηση της βάσης δεδομένων. Πρώτα, δίνεται μια λίστα με τις αλληλουχίες που είναι όμοιες με την αναζητούμενη αλληλουχία. Αυτή η λίστα περιέχει την ταυτότητα προέλευσης της βάσης δεδομένων (SW σημαίνει SWISS-PROT), το κωδικό όνομα (ID) της αλληλουχίας στη βάση δεδομένων, ένα κωδικό αριθμό (accession number), και τον τίτλο των όμοιων αλληλουχιών. Το μήκος των αμινοξέων από καθεμία ανακτόμενη όμοια αλληλουχία δηλώνεται σε αγκύλες. Μετά, δίνονται διάφορες βαθμολογίες που υπολογίστηκαν από το πρόγραμμα. Η πιο σημαντική βαθμολογία είναι η E-value, η οποία επιτρέπει την εκτίμηση της πιθανότητας για το εάν η ομοιότητα είναι πραγματική (το χαμηλό E-value δηλώνει ότι η ομοιότητα είναι πιθανώς πραγματική).

Μετά την σύνοψη της αναζήτησης, υπάρχουν τα αποτελέσματα από το σύνολο των συγκρίσεων ανά ζεύγη για συγκρίσεις με την ζητούμενη αλληλουχία που έχουν επιλεγεί εκ των προτέρων από τον χρήστη.

Μέσα στις αντιστοιχίες, οι ταυτίσεις δηλώνονται με τον χαρακτήρα ‘:’, και οι ομοιότητες με ‘.’ (οι ομοιότητες προσδιορίζονται με βάση ένα πίνακα ομοιότητας διαστάσεων 20x20 από αμινοξέα, τέτοιοι πίνακες είναι ο πίνακας Dayhoff και ο πίνακας BLOSUM).

Οι γραμμές από πάνω και από κάτω από τις αντιστοιχίες δηλώνουν τον απόλυτο αριθμό των αμινοξέων για τις σειρές (αγνοώντας τα κενά).

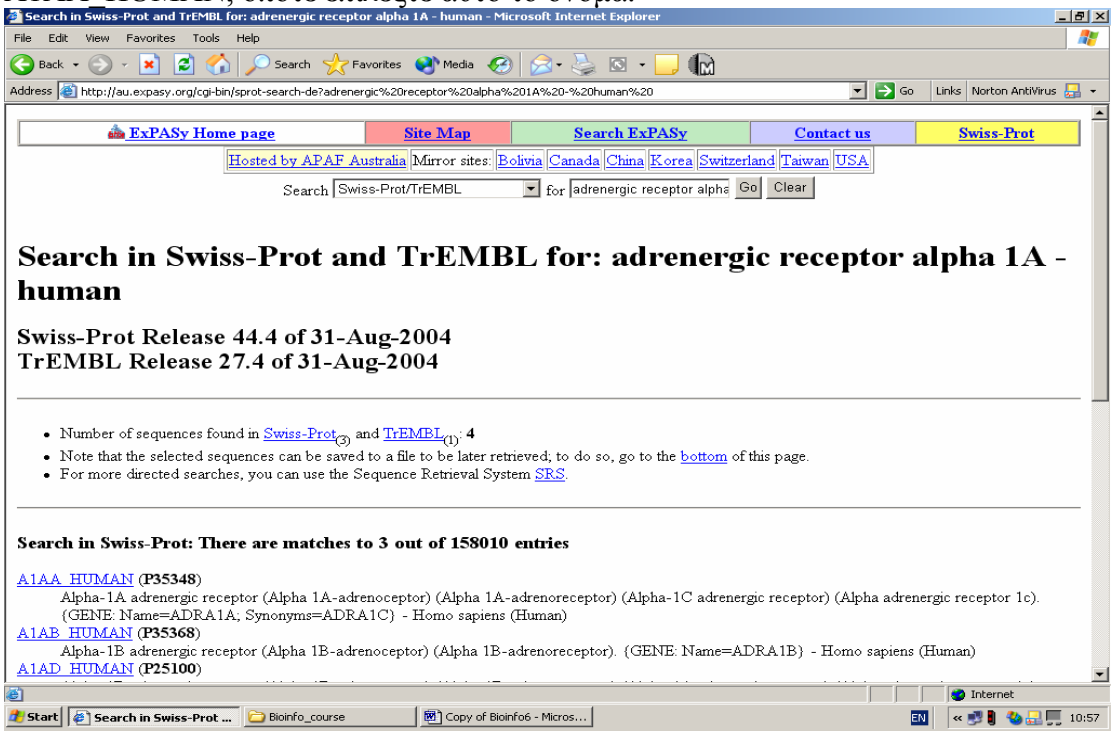
```
FASTA searches a protein or DNA sequence data bank
 version 3.4t23 March 18, 2004
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query library @ vs +swissprot library
searching /ebi/services/idata/v890/fastadb/swissprot library

  1>>>Sequence - 466 aa
 vs  SWISS-PROT Protein Sequence Database library

57769334 residues in 156998 sequences
 statistics extrapolated from 60000 to 155284 sequences
  Expectation_n fit: rho(ln(x))= 4.2996+/-0.000201; mu= 15.1018+/- 0.012
 mean_var=69.8740+/-14.547, 0's: 149 Z-trim: 758  B-trim: 8405 in 2/63
 Lambda= 0.153432

FASTA (3.47 Mar 2004) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 37, opt: 25, open/ext: -10/-2, width:  16
 Scan time: 24.750
The best scores are:                                      opt bits E(156998)
SW:A1AA_HUMAN P35348 Alpha-1A adrenergic receptor  ( 466) 3143 705.1 1.9e-202
SW:A1AA_RABIT O02824 Alpha-1A adrenergic receptor  ( 466) 2985 670.1 6.4e-192
SW:A1AA_CAVPO Q9WU25 Alpha-1A adrenergic receptor  ( 466) 2982 669.5  1e-191
SW:A1AA_BOVIN P18130 Alpha-1A adrenergic receptor  ( 466) 2944 661.0 3.5e-189
SW:A1AA_RAT P43140 Alpha-1A adrenergic receptor (  ( 466) 2941 660.4 5.5e-189
SW:A1AA_MOUSE P97718 Alpha-1A adrenergic receptor  ( 466) 2897 650.6 4.7e-186
SW:A1AA_CANFA O77621 Alpha-1A adrenergic receptor  ( 295) 1908 431.5 2.8e-120
SW:A1AA_ORYLA Q91175 Alpha-1A adrenergic receptor  ( 470) 1877 424.9 4.4e-118
…

>>SW:A1AA_RAT P43140 Alpha-1A adrenergic receptor (Alpha  (466 aa)
 initn: 2941 init1: 2941 opt: 2941  Z-score: 3518.7  bits: 660.4 E(): 5.5e-189
Smith-Waterman score: 2941;  92.704% identity (92.704% ungapped) in 466 aa
overlap (1-466:1-466)

               10        20        30        40        50        60
Sequen MVFLSGNASDSSNCTQPPAPVNISKAILLGVILGGLILFGVLGNILVILSVACHRHLHSV
       ::.:: ::::.::::.:::::::::::::::::::::.::::::::::::::::::::
SW:A1A MVLLSENASEGSNCTHPPAPVNISKAILLGVILGGLIIFGVLGNILVILSVACHRHLHSV
               10        20        30        40        50        60

               70        80        90       100       110       120
Sequen THYYIVNLAVADLLLTSTVLPFSAIFEVLGYWAFGRVFCNIWAAVDVLCCTASIMGLCII
       ::::::::::::::::::::::::::::.::::::::::::::::::::::::::::::::
SW:A1A THYYIVNLAVADLLLTSTVLPFSAIFEILGYWAFGRVFCNIWAAVDVLCCTASIMGLCII
               70        80        90       100       110       120

…
              370       380       390       400       410       420
Sequen HPPSQAVEGQHKDMVRIPVGSRETFYRISKTDGVCEWKFFSSMPRGSARITVSKDQSSCT
       ::::::.::::.::::::::: :::::::::::::::::::::::::: :::::.::
SW:A1A HPPSQALEGQHRDMVRIPVGSGETFYKISKTDGVCEWKFFSSMPQGSARITVPKDQSACT
              370       380       390       400       410       420

              430       440       450       460
Sequen TARVRSKSFLQVCCCVGPSTPSLDKNHQVPTIKVHTISLSENGEEV
       ::::::::::::::::: :.:  ..::::::::.::::.::::::
SW:A1A TARVRSKSFLQVCCCVGSSAPRPEENHQVPTIKIHTISLGENGEEV
       430     440     450     460
```

ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

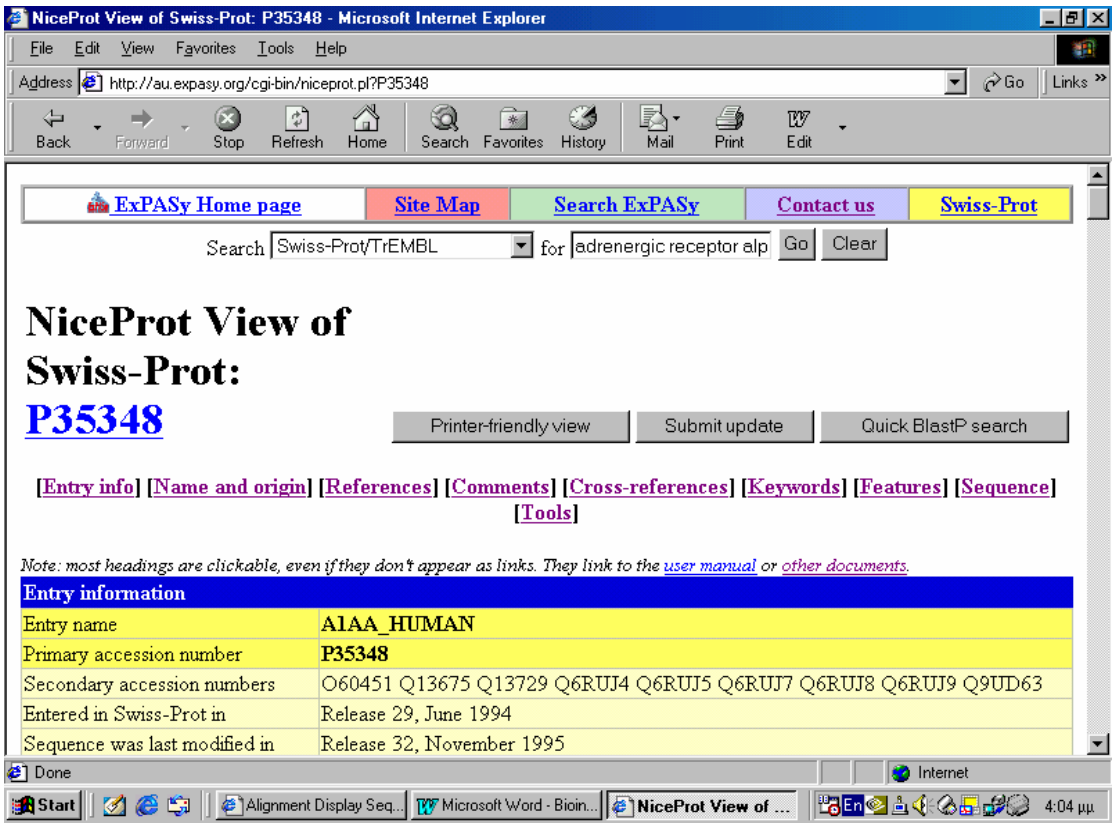
Αντιστοιχία αλληλουχιών ανά ζεύγη - FASTA

Ανοίξτε το Internet Explorer και πληκτρολογήστε τη διεύθυνση: au.expasy.org/sprot/sprot-top.html για την πρόσβαση στην βάση δεδομένων SWISS-PROT που περιέχει αλληλουχίες πρωτεϊνών. Μετά πληκτρολογήστε στο πλαίσιο παρακάτω: adrenergic receptor alpha 1A – human, και επιλέξτε "Go".

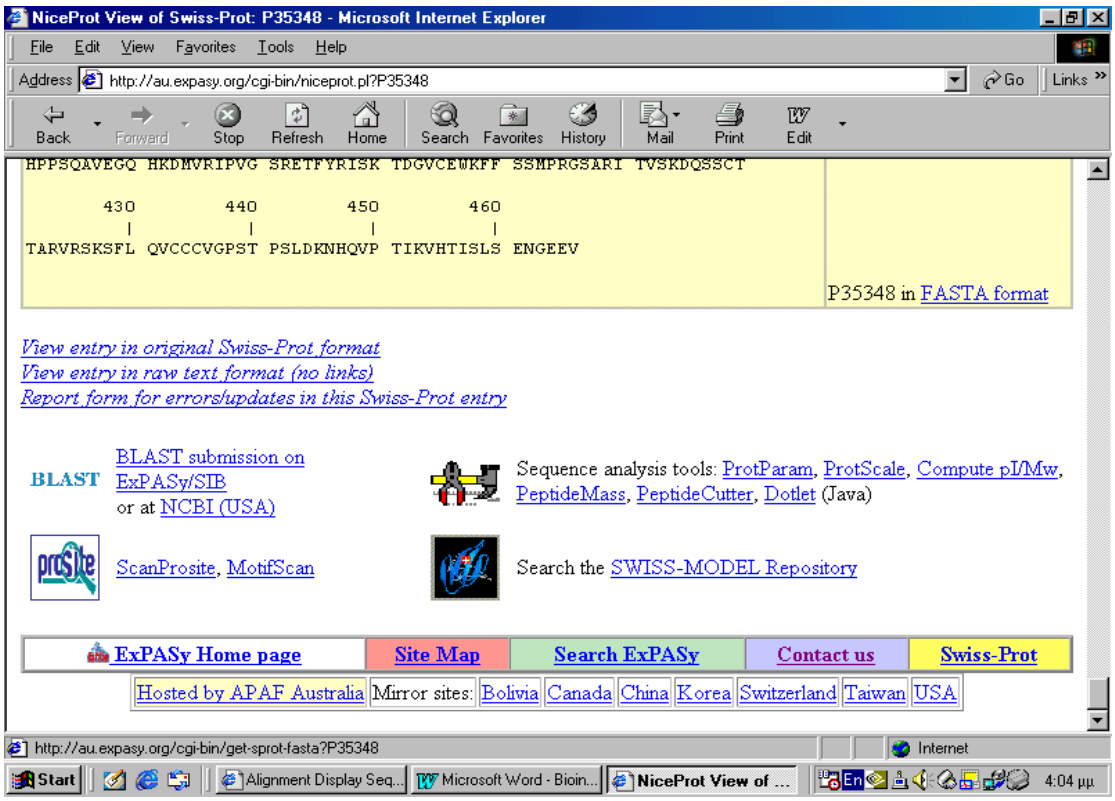
Στη συνέχεια, δίνεται μια λίστα σχετικών πρωτεϊνών. Η ζητούμενη πρωτεΐνη είναι η A1AA_HUMAN, οπότε επιλέξτε αυτό το όνομα.



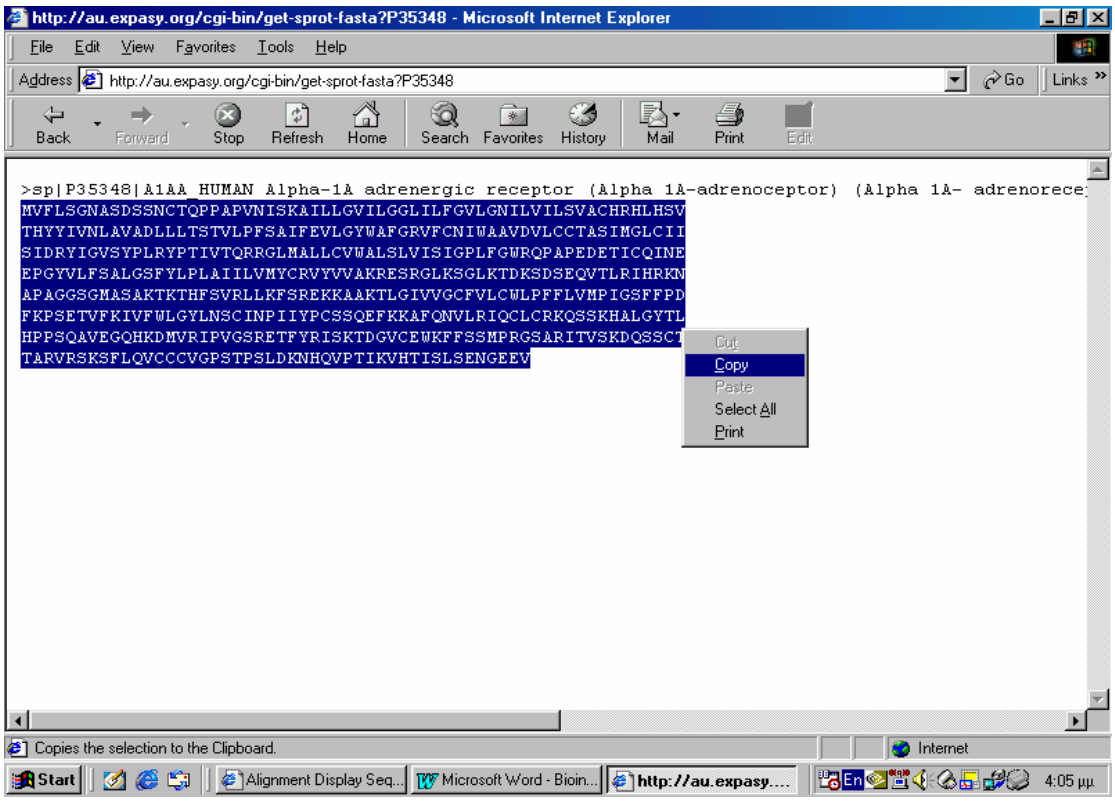
Μετά δίνονται οι πληροφορίες γι' αυτήν την πρωτεΐνη. Κυλήστε την οθόνη προς τα κάτω, στο τέλος της σελίδας.



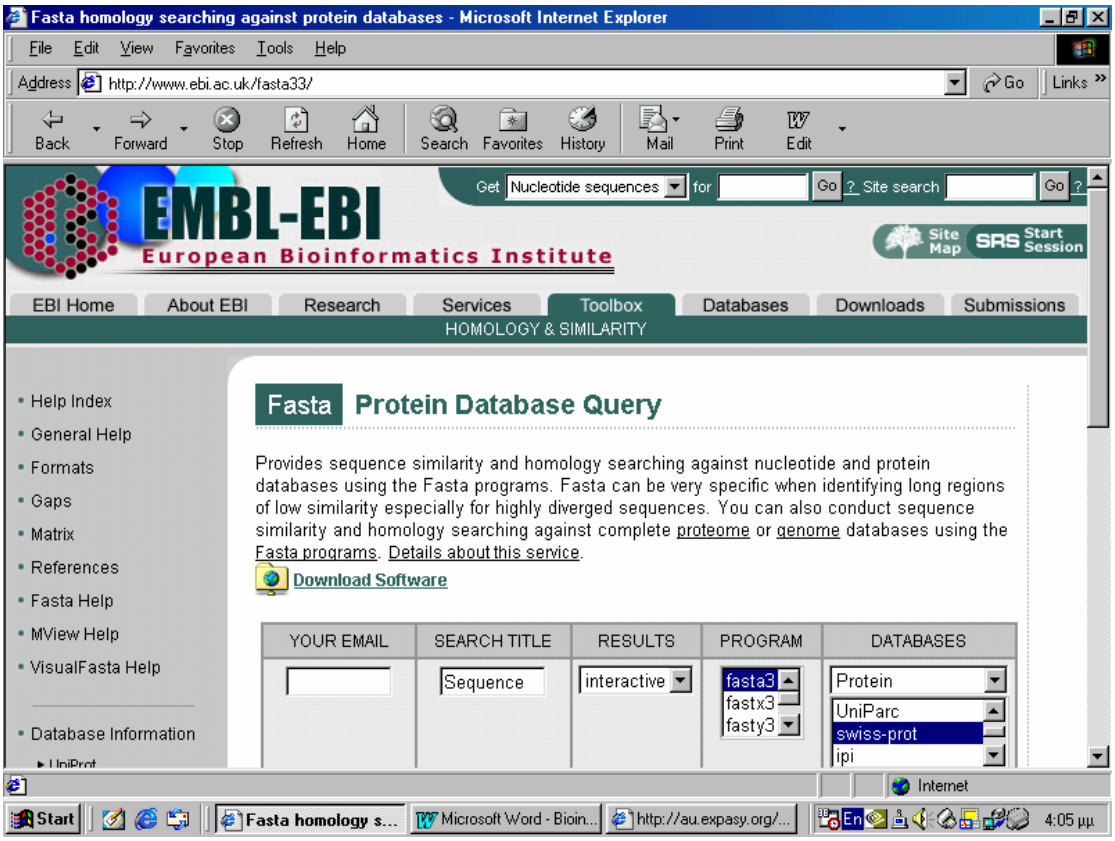
Επιλέξτε "FASTA format".



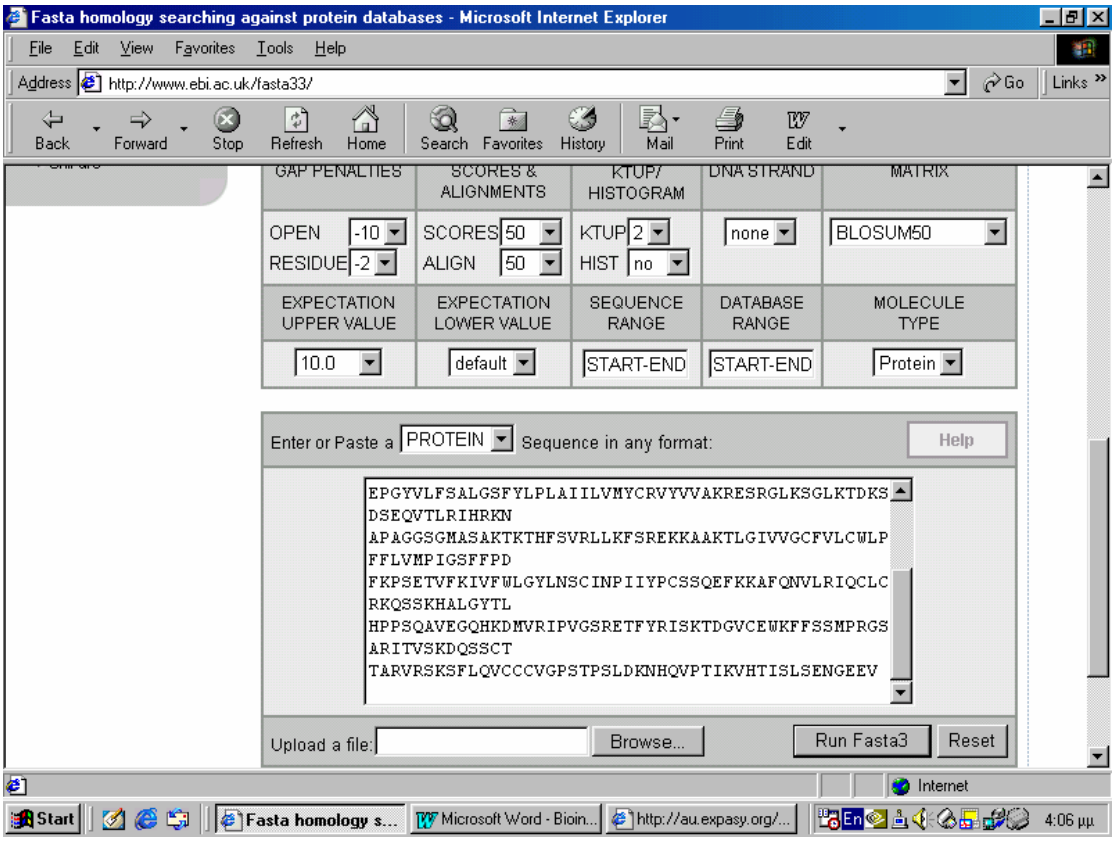
Τότε δίνεται η αλληλουχία της πρωτεΐνης, οπότε την επιλέγουμε και την αντιγράφουμε για να την επικολλήσουμε μέσα στο πρόγραμμα FASTA.



Για πρόσβαση στο FASTA, ανοίξτε το Internet Explorer και πληκτρολογήστε "www.ebi.ac.uk/fasta33/. Εκεί που γράφει "DATABASES" επιλέξτε "swiss-prot".

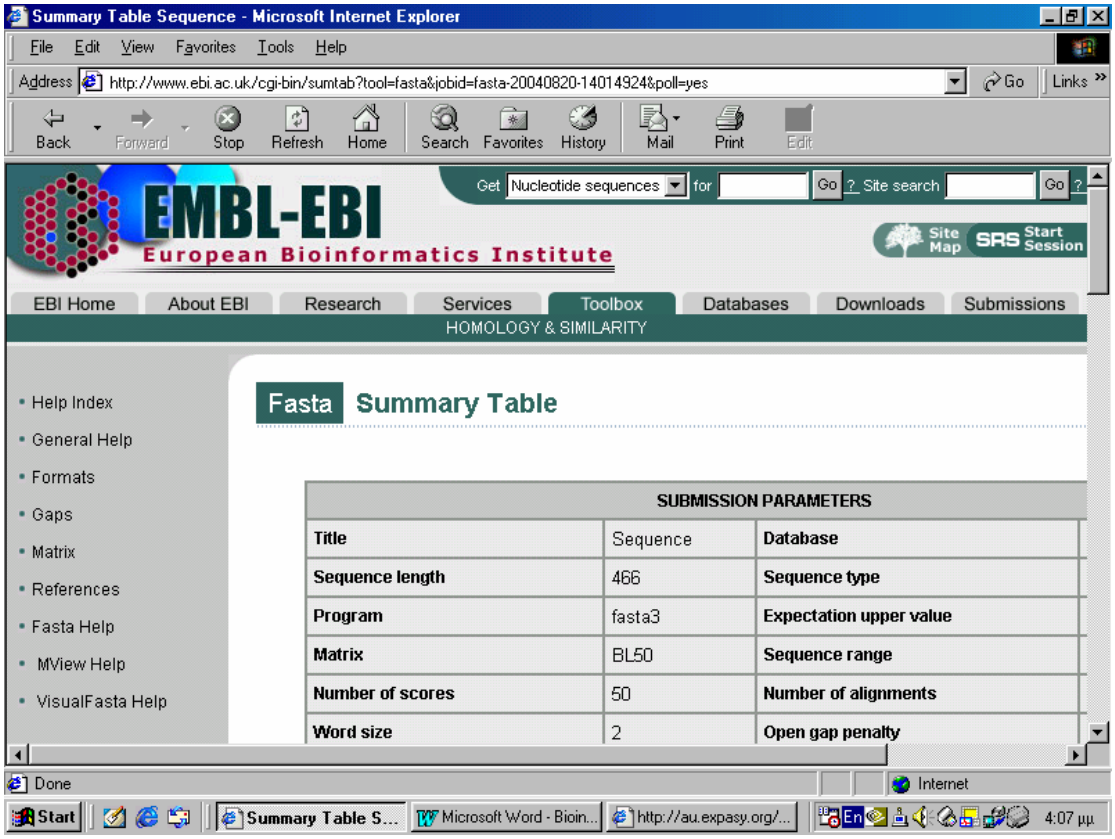


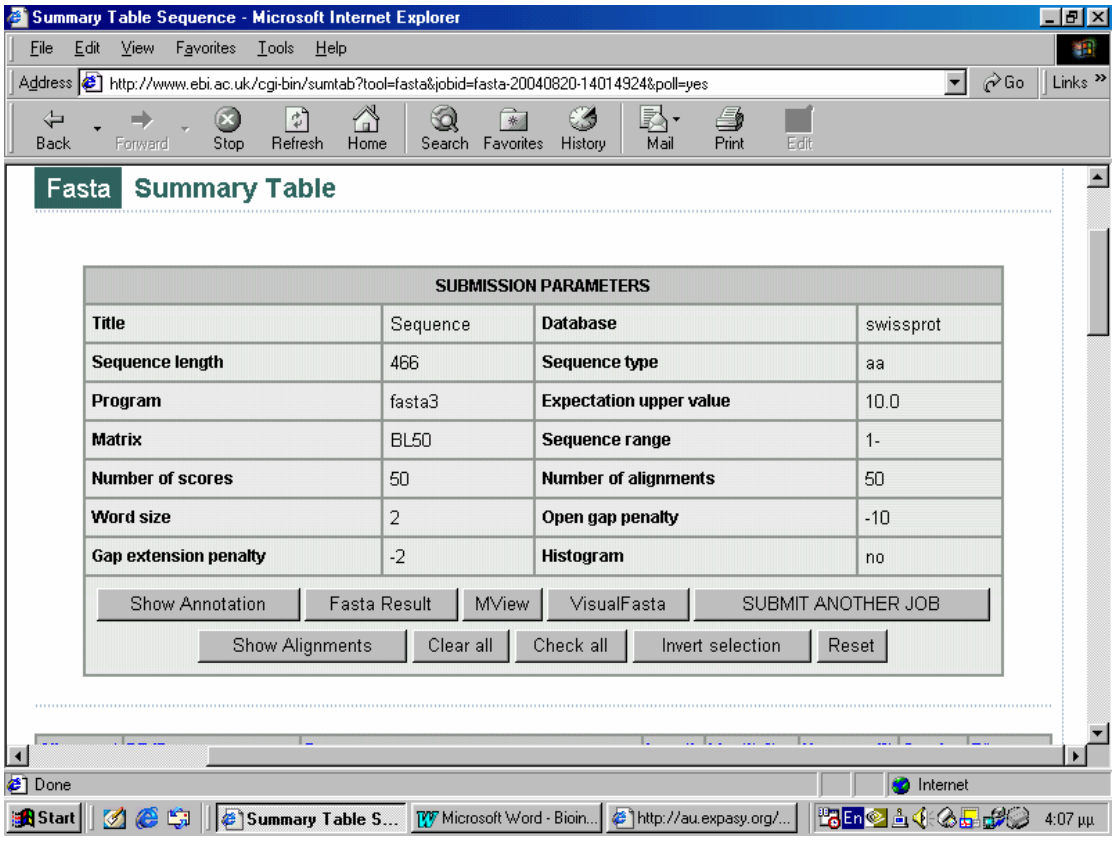
Στη συνέχεια, στο πλαίσιο παρακάτω επικολλήστε την αντιγραμμένη σειρά που θέλετε να



αντιστοιχίσετε. Μετά κάντε κλικ στο “Run Fasta3” και περιμένετε για μερικά δευτερόλεπτα.

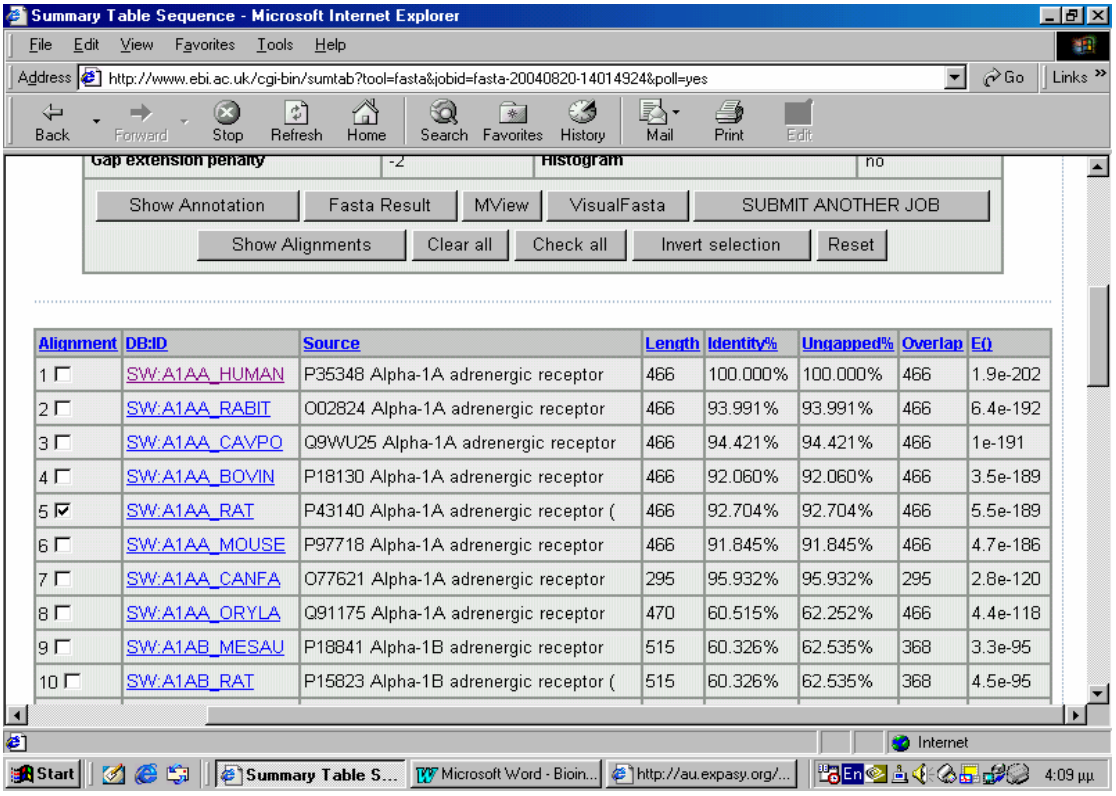
Μετά εμφανίζεται ένας συνοπτικός πίνακας διάφορες παραμέτρους.



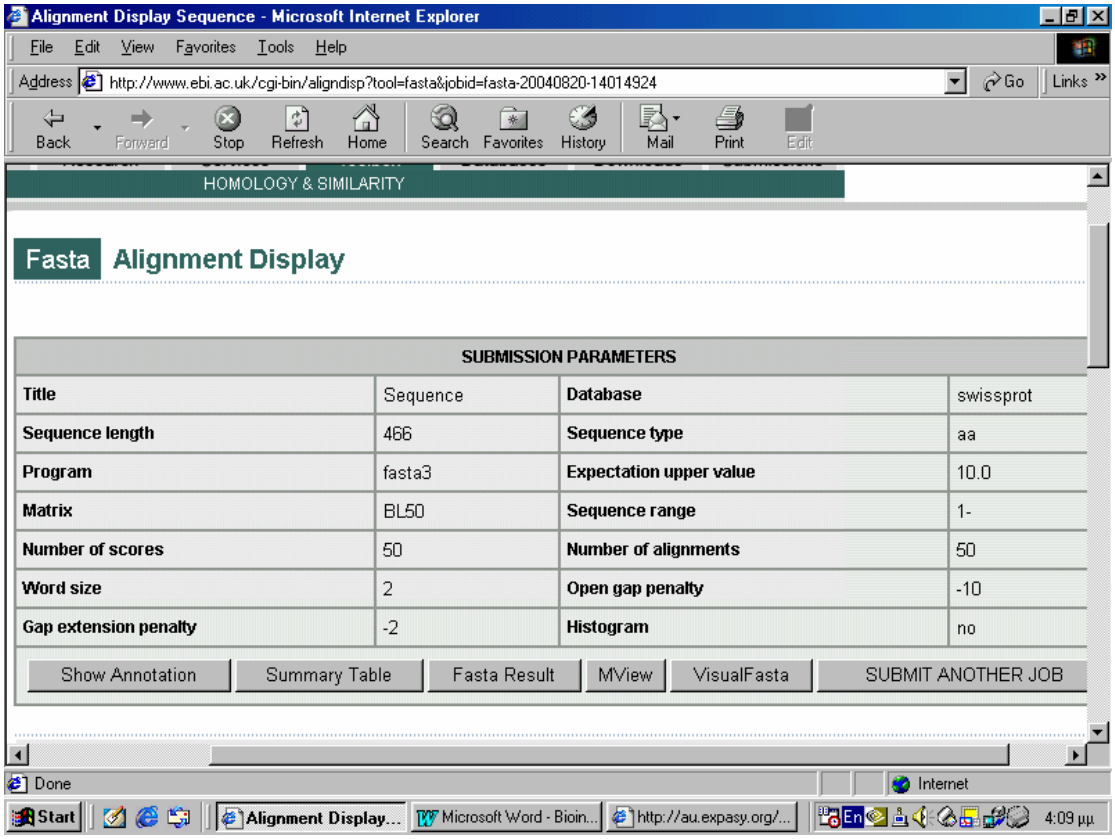


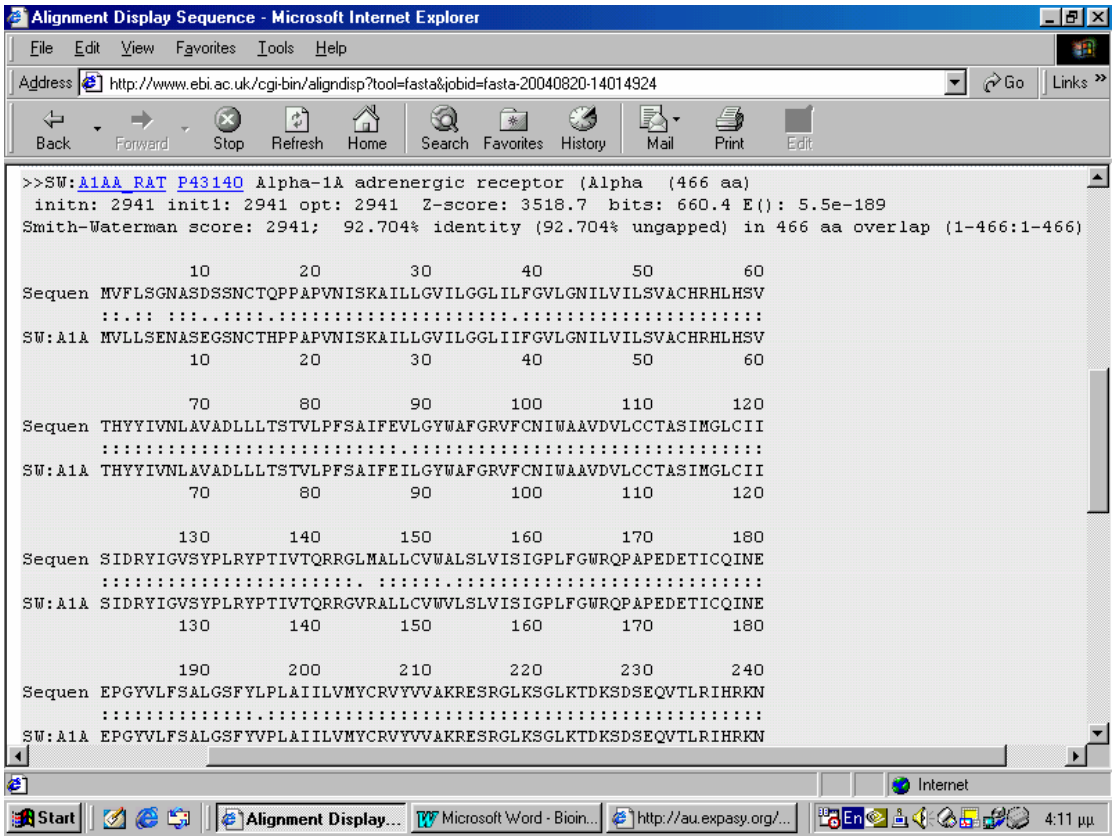
Τα αποτελέσματα (output) παρέχουν μια λίστα όλων των αλληλουχιών που είναι όμοιες με την ζητούμενη αλληλουχία. Οι αλληλουχίες προέρχονται από τη βάση δεδομένων SWISS-PROT. Το μήκος, ομοιότητα και το E-value για κάθε σύγκριση δίνεται. Αν ενδιαφέρεστε να

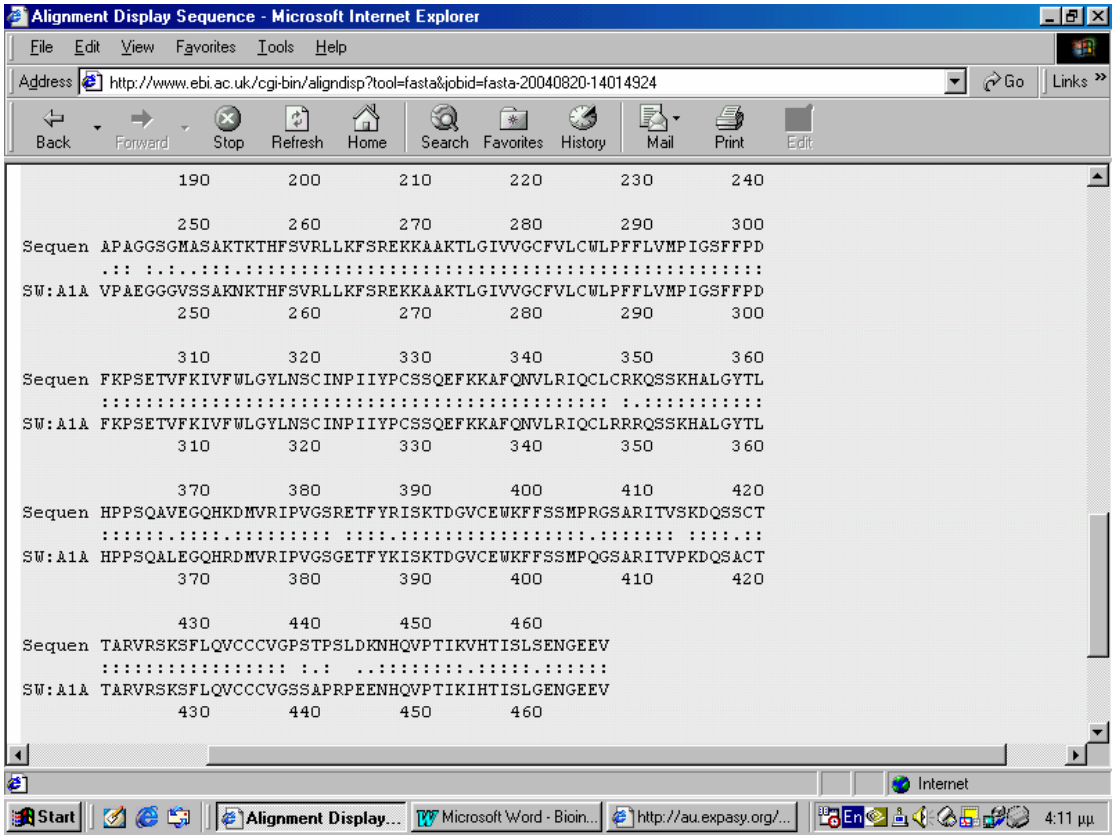
συγκρίνετε μόνο τις σειρές A1AA_HUMAN και A1AA_RAT, μην επιλέγεται (unmark) όλα τα πλαίσια κάτω από το “Alignment” εκτός από το A1AA_RAT. Μετά επιλέξτε το “Show Alignments”.



Τότε εμφανίζεται η αντιστοιχία.


Alignment Display Sequence - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.ebi.ac.uk/cgi-bin/alignndisp?tool=fasta&jobid=fasta-20040820-14014924
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
HOMOLOGY & SIMILARITY
Fasta Alignment Display
SUBMISSION PARAMETERS
Title Sequence Database swissprot
Sequence length 466 Sequence type aa
Program fasta3 Expectation upper value 10.0
Matrix BL50 Sequence range 1-
Number of scores 50 Number of alignments 50
Word size 2 Open gap penalty -10
Gap extension penalty -2 Histogram no
Show Annotation Summary Table Fasta Result MView VisualFasta SUBMIT ANOTHER JOB
Done
Internet
Start
Alignment Display...
Microsoft Word - Bioin...
http://au.expasy.org/...
4:09 μμ


Alignment Display Sequence - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.ebi.ac.uk/cgi-bin/alignndisp?tool=fasta&jobid=fasta-20040820-14014924
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
>>SW:A1AA_RAT P43140 Alpha-1A adrenergic receptor (Alpha (466 aa)
initn: 2941 init1: 2941 opt: 2941 Z-score: 3518.7 bits: 660.4 E(): 5.5e-189
Smith-Waterman score: 2941; 92.704% identity (92.704% ungapped) in 466 aa overlap (1-466:1-466)
10 20 30 40 50 60
Sequen MVFLSGNASDSSNCTQPPAPVNISKAILLGVILGGLILFGVLGNILVILSVACHRHLHSV
::.:: ::::.:::::.::::::::::::::::::.::::::::::::::::::::
SW:A1A MVLLSENASEGSNCTHPPAPVNISKAILLGVILGGLIIFGVLGNILVILSVACHRHLHSV
10 20 30 40 50 60
70 80 90 100 110 120
Sequen THYYIVNLAVADLLLTSTVLPFSAIFEVLGYWAFGRVFCNIWAAVDVLCCTASIMGLCII
::::::::::::::::::::::::::.::::::::::::::::::::::::::::::
SW:A1A THYYIVNLAVADLLLTSTVLPFSAIFEILGYWAFGRVFCNIWAAVDVLCCTASIMGLCII
70 80 90 100 110 120
130 140 150 160 170 180
Sequen SIDRYIGVSYPLRYPTIVTQRRGLMALLCVWALSLVISIGPLFGWRQPAPEDETICQINE
:::::::::::::::::::::: :::::::::::::::::::::::::::::::::
SW:A1A SIDRYIGVSYPLRYPTIVTQRRGVRALLCVWVLSLVISIGPLFGWRQPAPEDETICQINE
130 140 150 160 170 180
190 200 210 220 230 240
Sequen EPGYVLFSALGSFYLPLAIILVMYCRVYVVAKRESRGLKSGLKTDKSDSEQVTLRIHRKN
::::::::::::::.::
SW:A1A EPGYVLFSALGSFYVPLAIILVMYCRVYVVAKRESRGLKSGLKTDKSDSEQVTLRIHRKN
Internet
Start
Alignment Display...
Microsoft Word - Bioin...
http://au.expasy.org/...
4:11 μμ


Alignment Display Sequence - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.ebi.ac.uk/cgi-bin/aligndisp?tool=fasta&jobid=fasta-20040820-14014924
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
190 200 210 220 230 240
250 260 270 280 290 300
Sequen APAGGSGMASAKTKTHFSVRLLKFSREKKAAKTLGIVVGCFVLCWLPFFLVMPIGSFFPD
.:: :.:..:::.::
SW:A1A VPAEGGGVSSAKNKTHFSVRLLKFSREKKAAKTLGIVVGCFVLCWLPFFLVMPIGSFFPD
250 260 270 280 290 300
310 320 330 340 350 360
Sequen FKPSETVFKIVFWLGYLNSCINPIIYPCSSQEFKKAFQNVLRIQCLCRKQSSKHALGYTL
:: :.::::::::::::
SW:A1A FKPSETVFKIVFWLGYLNSCINPIIYPCSSQEFKKAFQNVLRIQCLRRRQSSKHALGYTL
310 320 330 340 350 360
370 380 390 400 410 420
Sequen HPPSQAVEGQHKDMVRIPVGSRETFYRISKTDGVCEWKFFSSMPRGSARITVSKDQSSCT
::::::.::::.::::::::: ::::.::::::::::::::::.::::::: ::::.::
SW:A1A HPPSQALEGQHRDMVRIPVGSGETFYKISKTDGVCEWKFFSSMPQGSARITVPKDQSACT
370 380 390 400 410 420
430 440 450 460
Sequen TARVRSKSFLQVCCCVGPSTPSLDKNHQVPTIKVHTISLSENGEEV
::::::::::::::::: :.: ..:::::::::.::::::.::::::
SW:A1A TARVRSKSFLQVCCCVGSSAPRPEENHQVPTIKIHTISLGENGEEV
430 440 450 460
Internet
Start Alignment Display... Microsoft Word - Bioin... http://au.expasy.org/... 4:11 µµ

Πολλαπλή αντιστοιχία αλληλουχίας

Η πολλαπλή αντιστοιχία απεικονίζει τις σχέσεις μεταξύ δύο ή περισσοτέρων αλληλουχιών. Όταν οι αλληλουχίες που εμπλέκονται διαφέρουν, τα διατηρημένα (conserved) αμινοξέα και στις δύο αλληλουχίες είναι συνήθως σχετίζονται με τη διατήρηση της σταθερότητας της δομής και της βιολογικής λειτουργίας της πρωτεΐνης.
Αν ένα αμινοξύ διατηρείται σε μία οικογένεια αλληλουχιών, οι οποίες προέρχονται από διαφορετικά είδη, τότε αυτό δείχνει ότι το αμινοξύ παίζει ένα σημαντικό ρόλο στη δομή ή λειτουργία της πρωτεΐνης. Ένα τέτοιο αμινοξύ μπορεί να αναγνωριστεί με πολλαπλή αντιστοιχία.

Ορισμός

Η πολλαπλή αντιστοιχία αλληλουχιών είναι ένας 2D πίνακας, όπου οι γραμμές αντιπροσωπεύουν τις διαφορετικές αλληλουχίες και οι στήλες τις θέσεις των αμινοξέων. Οι αλληλουχίες τοποθετούνται στο πίνακα έτσι ώστε α) τα πιο όμοια αμινοξέα να περιλαμβάνονται σε κάθετη εγγραφή, με τη χρήση κενών (-) και β) η τάξη των αμινοξέων σε κάθε αλληλουχία να διατηρείται. Για παράδειγμα, μια πολλαπλή αντιστοιχία για πέντε είδη αλληλουχιών (I-V) είναι η παρακάτω:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
I	Y	D	G	G	A	V	-	E	A	L
II	Y	D	G	G	-	-	-	E	A	L
III	F	E	G	G	I	L	V	E	A	L
IV	F	D	-	G	I	L	V	Q	A	V
V	Y	E	G	G	A	V	V	Q	A	L

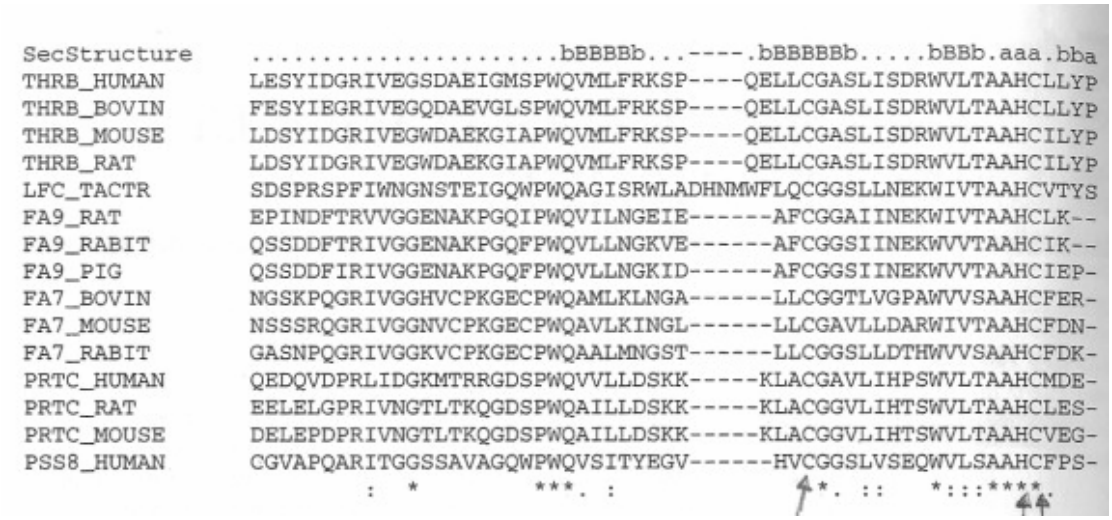
Κοινή-συνοπτική αλληλουχία

Ο πίνακας αντιστοίχισης μπορεί να συνοψισθεί σε μια γραμμή (ψευδό-αλληλουχία [pseudo-sequence]) που προσθέτεται στο τέλος της αντιστοιχίας. Αυτή η ψευδό-αλληλουχία αποτελείται από σύμβολα, τα οποία συνοψίζουν τον χαρακτήρα της αντιστοιχίας σε κάθε κάθετη θέση, με τον παρακάτω τρόπο: 1) αν υπάρχει μόνο ένα σύμβολο αμινοξέως, τότε χρησιμοποιείται κεφαλαίο γράμμα, 2) αν η πλειοψηφία των συμβόλων είναι ένα γράμμα τότε χρησιμοποιείται μικρό γράμμα, 3) αν υπάρχει ίσος αριθμός διαφορετικών αμινοξέων τότε χρησιμοποιούνται όλα τα αμινοξέα.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
I	Y	D	G	G	A	V	-	E	A	L
II	Y	D	G	G	-	-	-	E	A	L
III	F	E	G	G	I	L	V	E	A	L
IV	F	D	-	G	I	L	V	Q	A	V
V	Y	E	G	G	A	V	V	Q	A	L
	Y	**d**	**G**	**G**	**A/I**	**V/L**	**V**	**e**	**A**	**l**

Διατηρημένα αμινοξέα – Παράδειγμα

Ένα παράδειγμα πολλαπλής αντιστοιχίας μερικών αλληλουχιών της serine protease παρουσιάζεται παρακάτω. Αυτή η αντιστοιχία δείχνει ότι υπάρχον δύο κύριοι λόγοι για την διατήρηση αμινοξέων στις πρωτεΐνες: 1) για τη διατήρηση της λειτουργίας και 2) για τη διατήρηση της δομής.



Η ιστιδίνη (H) είναι απαραίτητη για τη λειτουργία αυτών των ενζύμων (πρωτεϊνών). Επειδή είναι απαραίτητη στη λειτουργία, διατηρείται σε κάθε μέλος της οικογενείας. Η διατηρημένη ιστιδίνη είναι τοποθετημένη στην $6^{η}$ θέση από τη δεξιά μεριά της αντιστοιχίας και όπως αναμένεται διατηρείται σε όλες τις αντιστοιχισμένες αλληλουχίες.
Η κυστεϊνη (C) είναι απαραίτητη για τη διατήρηση μιας σταθερής τρισδιάστατης (3D) δομής. Τα βασικά διατηρημένα στοιχεία της δομής είναι οι δύο διατηρημένες κυστεϊνες στην αντιστοιχία, οι οποίες είναι τοποθετημένες στην $5^{η}$ και $20^{η}$ θέση της αντιστοιχίας και διατηρούνται σε όλες τις αντιστοιχισμένες αλληλουχίες (δηλ. αυτές οι δυο συγκεκριμένες κυστεϊνες σε αυτές τις θέσεις είναι υπεύθυνες για τη διατήρηση της δομής της πρωτεΐνης).

Υπολογισμός – τεχνικές

Οι τεχνικές πολλαπλής αντιστοιχίας αλληλουχίας χρειάζονται χρόνο επεξεργασίας και χώρο στη μνήμη ανάλογα με το μέγεθος των αλληλουχιών που συγκρίνονται: O(m1m2m3…ml), όπου το O είναι η τάξη του χρόνου που χρειάζεται ο αλγόριθμος και mi είναι το μήκος της αλληλουχίας i. Οπότε, ο χρόνος που χρειάζεται για να υπολογιστεί μια αντιστοιχία αυξάνει εκθετικά με όσες πιο πολλές αλληλουχίες αντιστοιχίζονται.
Έχουν δημιουργηθεί διάφορες τεχνικές για να μειωθεί ο χρόνος για την εύρεση καλών αντιστοιχιών, αυτές οι τεχνικές συμπεριλαμβάνουν: 1) αντιστοίχιση όλων των αλληλουχιών ανά ζεύγη, 2) αντιστοίχιση κάθε αλληλουχίας με μια συγκεκριμένη αλληλουχία, αντιστοίχιση αλληλουχιών σε αυθαίρετη τάξη, ή 3) αντιστοίχιση αλληλουχιών ακολουθώντας την τάξη διακλάδωσης ενός φυλογενετικού δέντρου.

Πρακτική αντιστοίχιση

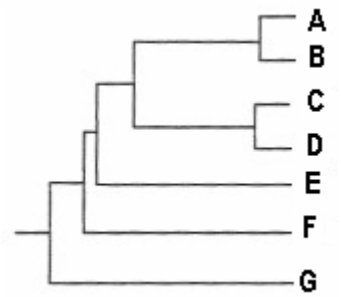
Η πολλαπλή αντιστοίχιση αλληλουχίας μπορεί να γίνει με τη χρήση πρακτικών μεθόδων. Τα πακέτα ανάλυσης αλληλουχιών προσφέρουν και οθόνες (editors) πρακτικής αντιστοίχισης. Για τον οπτικό εντοπισμό ομοιοτήτων μεταξύ των αλληλουχιών χρησιμοποιούνται διάφορα χρώματα. Η κατάλληλη επιλογή χρώματος προσφέρει ένα γρήγορο τρόπο για την αποκάλυψη των περιοχών που είναι σημαντικές για τη δομή ή τη λειτουργία μιας πρωτεΐνης (δηλ. μπορούν να απεικονισθούν εύκολα σημαντικά διατηρημένα αμινοξέα, ή ομάδες αμινοξέων, και ασυνήθιστες μεταλλάξεις). Στα πακέτα 3D γραφικών για πρωτεΐνες χρησιμοποιούνται τα παρακάτω χρώματα:

Residue	**Property**	**Color**
Asp, Glu	Acidic	red
His, Arg, Lys	Basic	blue
Ser, Thr, Asn, Gln	Polar neutral	green
Ala, Val, Leu, Ile, Met	Hydrophobic aliphatic	white
Phe, Tyr, Trp	Hydrophobic aromatic	purple
Pro, Gly	Special structural properties	brown
Cys	Disulphide bond former	yellow

Η ποιοτική εκτίμηση της σχέσης όλων των ζευγών αλληλουχιών μπορεί να υπολογιστεί με τον υπολογισμό των ταυτόσημων και ομοιοτήτων αμινοξέων.

Προοδευτική αντιστοιχία

Η προοδευτική αντιστοιχία, πρώτα συμπεριλαμβάνει μια αρχική εκτίμηση για το πώς οι αλληλουχίες σχετίζονται χρησιμοποιώντας αντιστοιχίες ανά ζεύγη, και μετά δημιουργείται ένα καθοδηγητικό δέντρο. Με τη χρήση αυτού του καθοδηγητικού δέντρου, προσθέτονται αλληλουχίες προοδευτικά στην αντιστοιχία, ξεκινώντας με τις πιο σχετιζόμενες αλληλουχίες και τελειώνοντας με τις πιο απομακρυσμένες. Η διαδικασία είναι όπως παρακάτω:

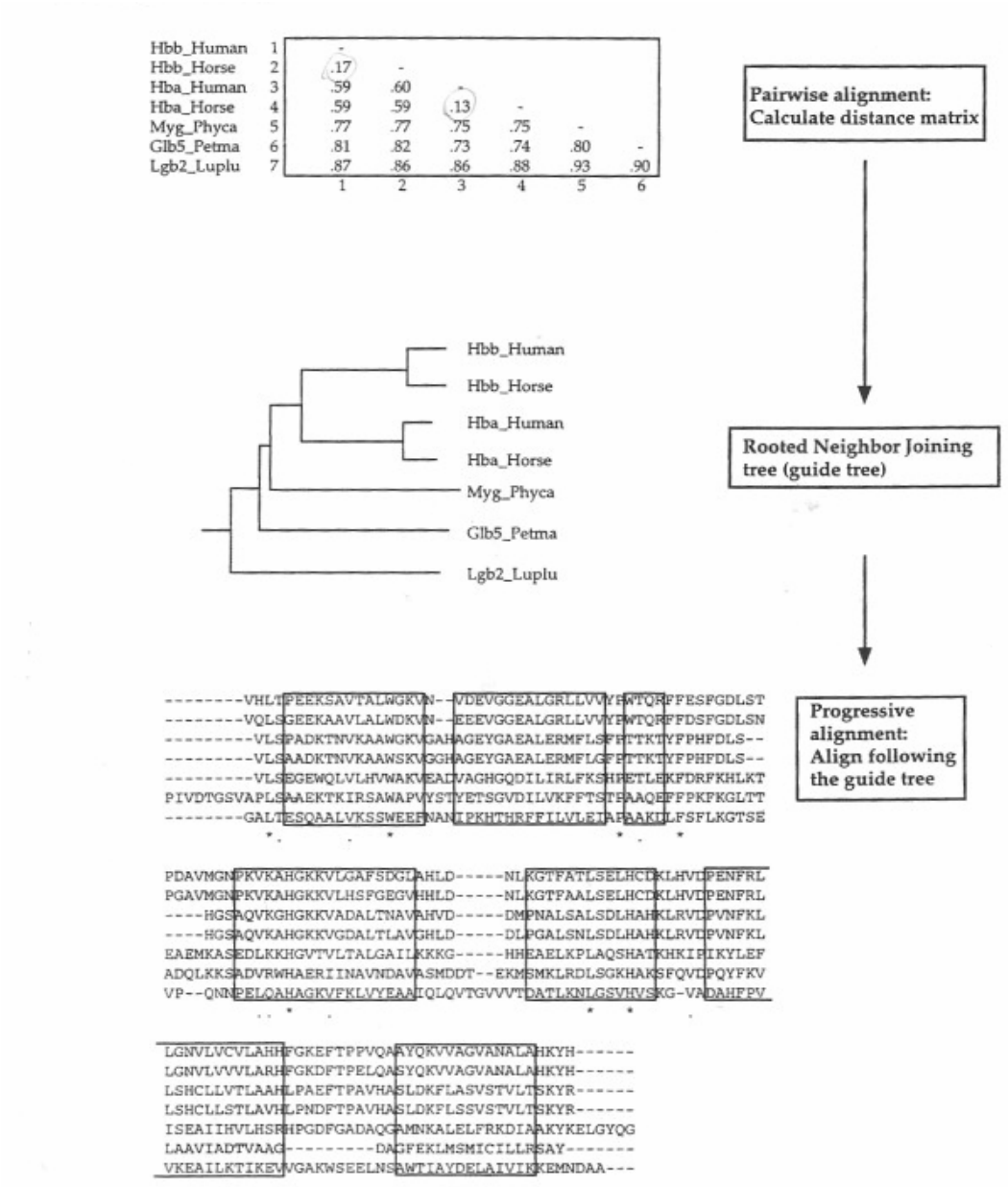


Αντίστοιχη αλληλουχία αντιστοιχίας
1. A με B ===> αντιστοιχία AB
2. C με D ===> αντιστοιχία CD
3. Αντιστοιχία AB με αντιστοιχία CD ===> ABCD
4. ABCD με E ===> ABCDE
5. ABCDE με F ===> ABCDEF
6. ABCDEF με G ===> ABCDEFG

Αυτό το δέντρο μπορεί να θεωρηθεί ως ένας πιθανός τρόπος με τον οποίο οι αλληλουχίες εξελίχτηκαν (φυλογενετικό δέντρο- phylogenetic tree). Η A και B είναι πολύ σχετιζόμενες και έχουν απομακρυνθεί από έναν κοντινό κοινό πρόγονο, όπως συμβαίνει και με την C και D. Αυτές οι 4 αντιστοιχίες έχουν ένα κοινό πρόγονο και είναι πιο πολύ σχετιζόμενες μεταξύ τους απ' ότι με τις αλληλουχίες E, F και G. Οπότε η αλληλουχία της αντιστοιχίας είναι η ακόλουθη: Πρώτα, αντιστοιχίζονται οι πιο σχετιζόμενες αλληλουχίες A και B, ακολουθούν οι αλληλουχίες C και D και μετά αυτές οι δύο αντιστοιχίες αντιστοιχίζονται για να δημιουργηθεί μια αντιστοιχία από τέσσερις αλληλουχίες. Στη συνέχεια, οι αλληλουχίες E, F και G προσθέτονται διαδοχικά σ' αυτήν την αντιστοιχία. Όλες αυτές οι αντιστοιχίες πραγματοποιούνται με το δυναμικό προγραμματισμό.

ClustalW

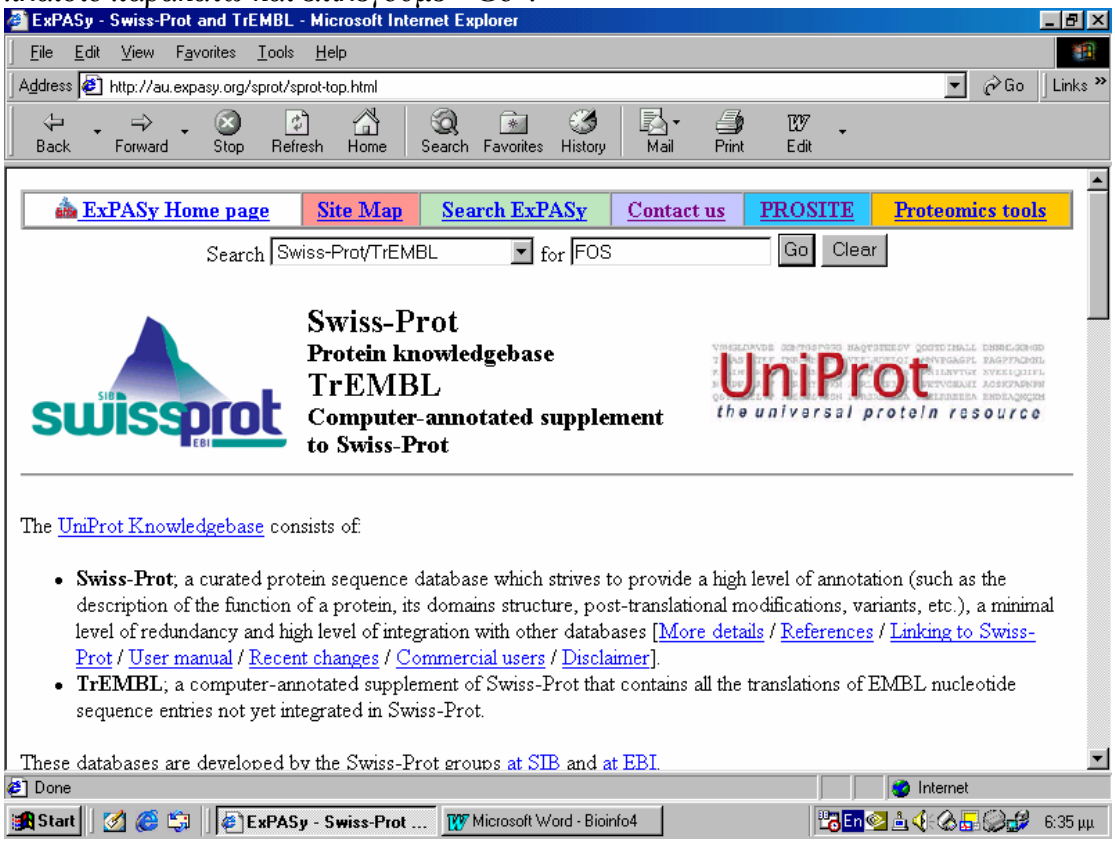
Το ClustalW είναι το πιο γνωστό πρόγραμμα υπολογιστή για την εκτέλεση μιας προοδευτικής αντιστοιχίας. Σ' αυτό το πρόγραμμα, η θέση των κενών στις πιο σχετιζόμενες αλληλουχίες χρησιμοποιείται για την καθοδήγηση της εισαγωγής κενών μέσα στις αλληλουχίες που είναι πιο απομακρυσμένες. Παρακάτω, παρουσιάζεται η πολλαπλή αντιστοίχιση επτά αλληλουχιών από globins, χρησιμοποιώντας το πρόγραμμα ClustalW. Η ανάκτηση των αλληλουχιών έγινε από την SWISS-PROT, τη βάση δεδομένων με πρωτεΐνες.


Hbb_Human 1 -
Hbb_Horse 2 .17 -
Hba_Human 3 .59 .60 -
Hba_Horse 4 .59 .59 .13 -
Myg_Phyca 5 .77 .77 .75 .75 -
Glb5_Petma 6 .81 .82 .73 .74 .80 -
Lgb2_Luplu 7 .87 .86 .86 .88 .93 .90
1 2 3 4 5 6
Pairwise alignment: Calculate distance matrix
Hbb_Human
Hbb_Horse
Hba_Human
Hba_Horse
Myg_Phyca
Glb5_Petma
Lgb2_Luplu
Rooted Neighbor Joining tree (guide tree)
Progressive alignment: Align following the guide tree

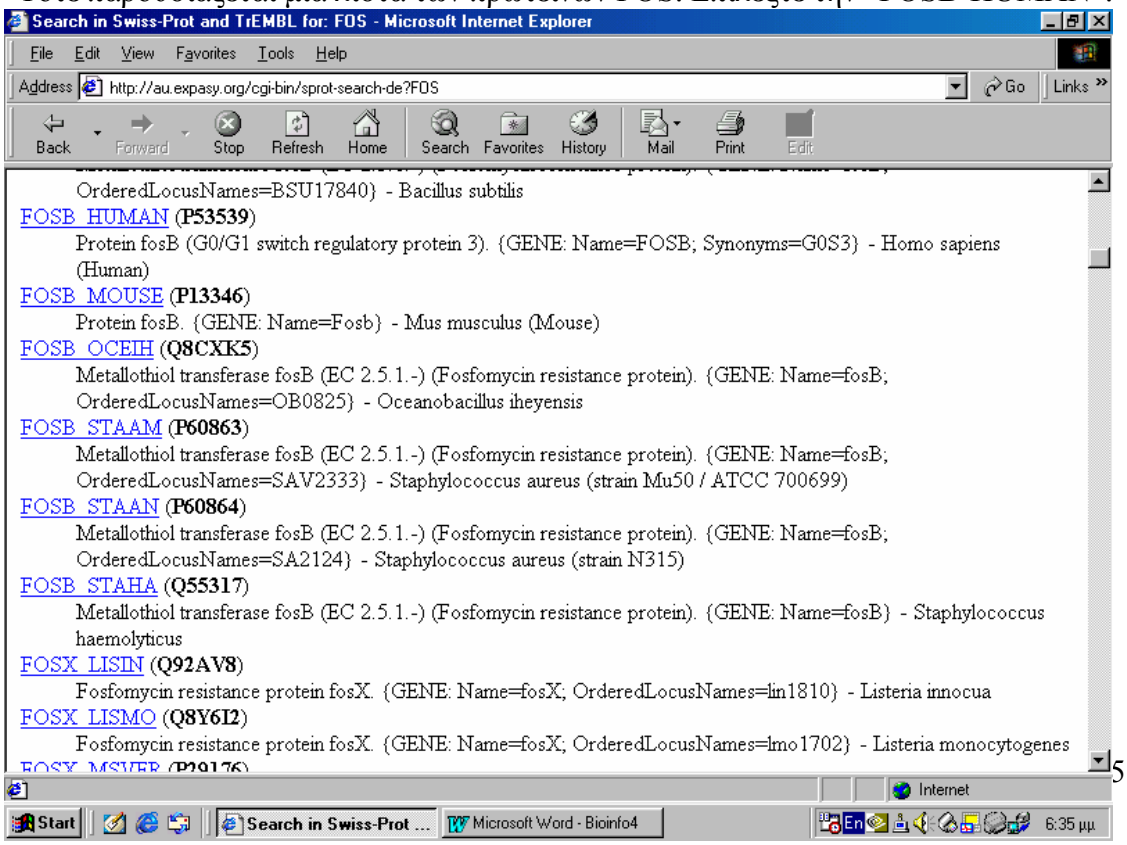
ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

Πολλαπλή αντιστοίχιση αλληλουχίας – ClustalW

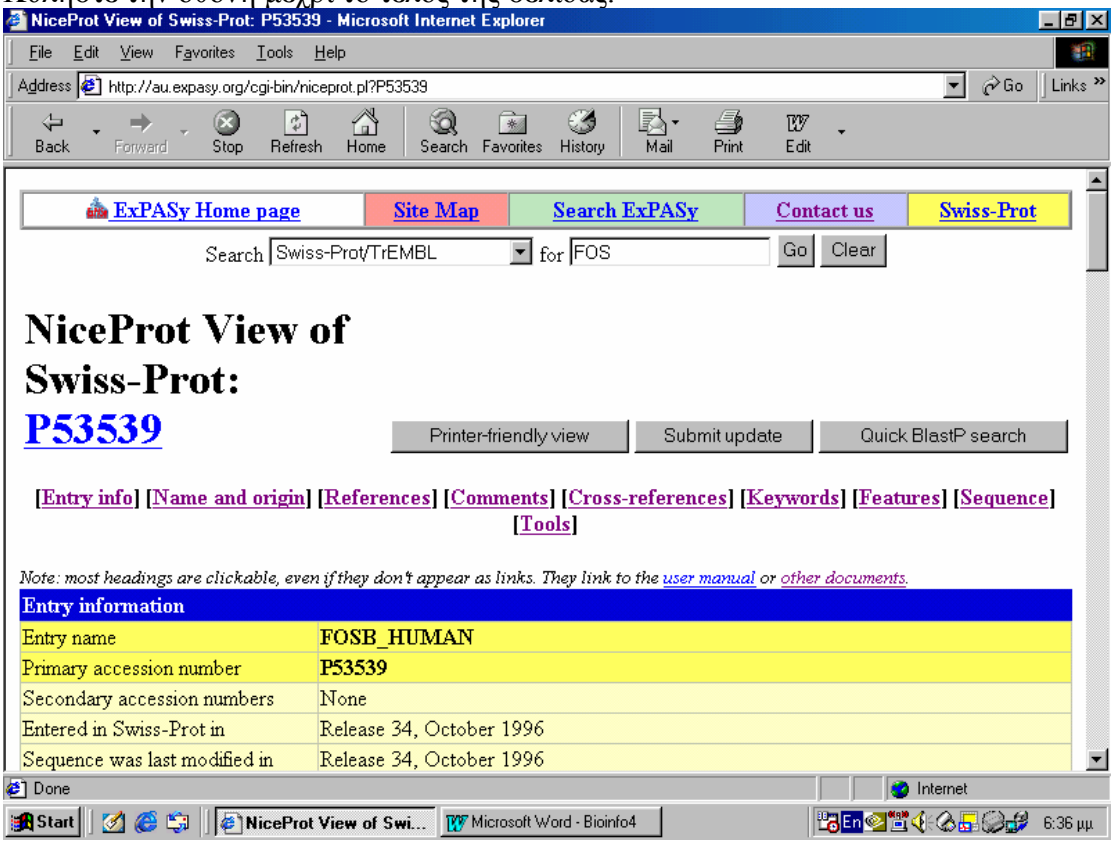
Θέλουμε να αντιστοιχίσουμε 5 πρωτεΐνες FOS από διαφορετικά είδη. Για να ανακτήσουμε τις αλληλουχίες των πρωτεϊνών, πρέπει να μπούμε στη βάση δεδομένων SWISS-PROT, οπότε ανοίγουμε το Internet Explorer και πληκτρολογούμε τη διεύθυνση: au.expasy.org/sprot/sprot-top.html. Μετά πληκτρολογήστε FOS στο πλαίσιο παρακάτω και επιλέγουμε "Go".



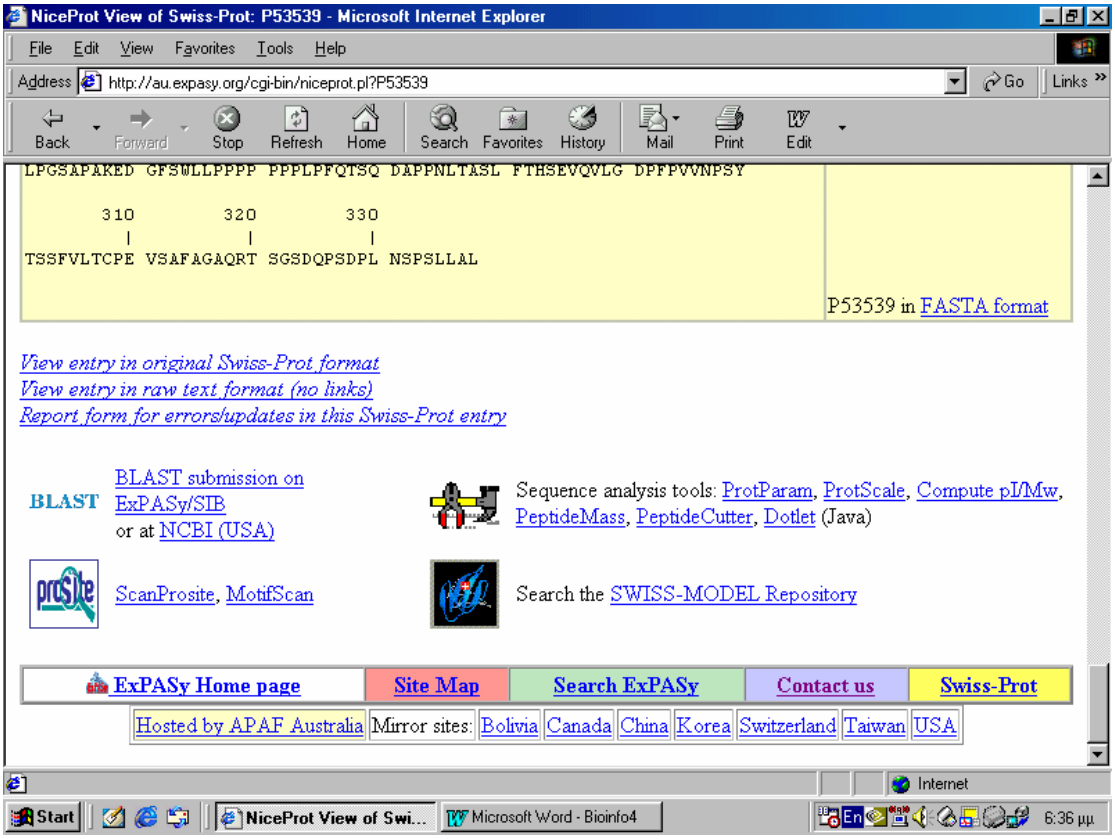
Τότε παρουσιάζεται μια λίστα των πρωτεϊνών FOS. Επιλέξτε την "FOSB-HUMAN".



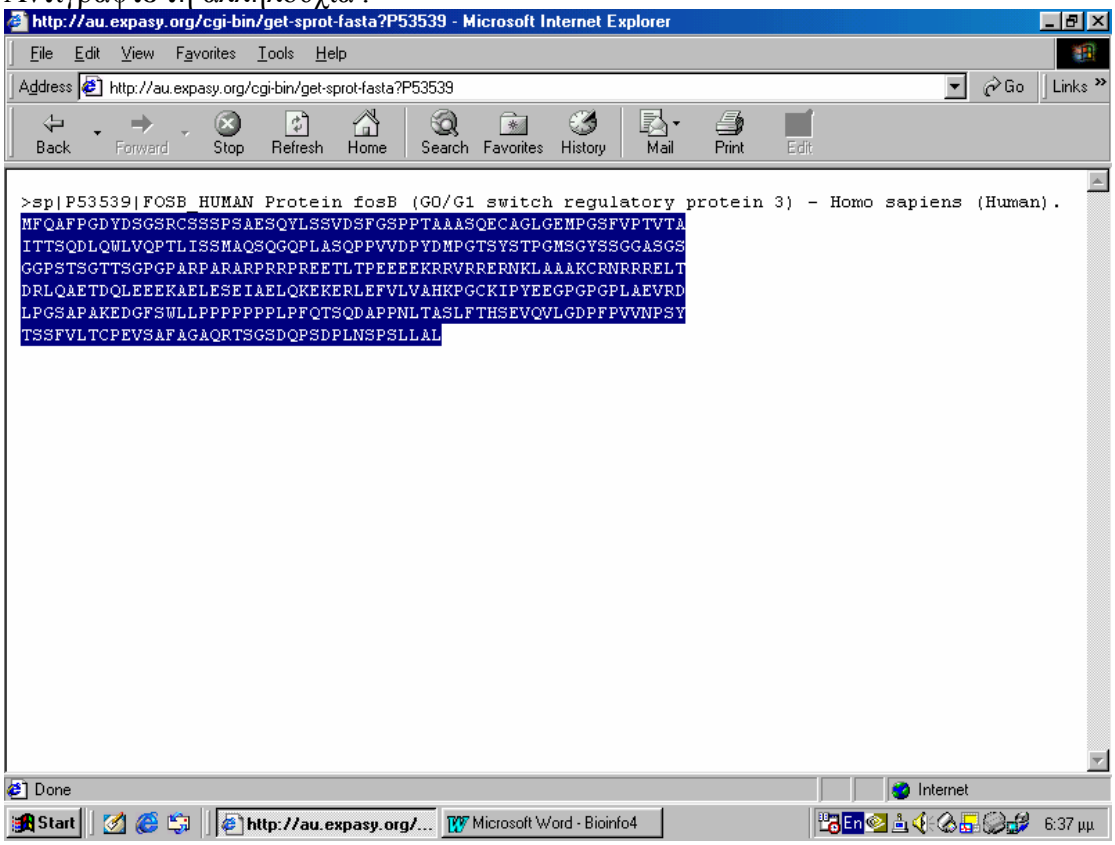
Κυλήστε την οθόνη μέχρι το τέλος της σελίδας.



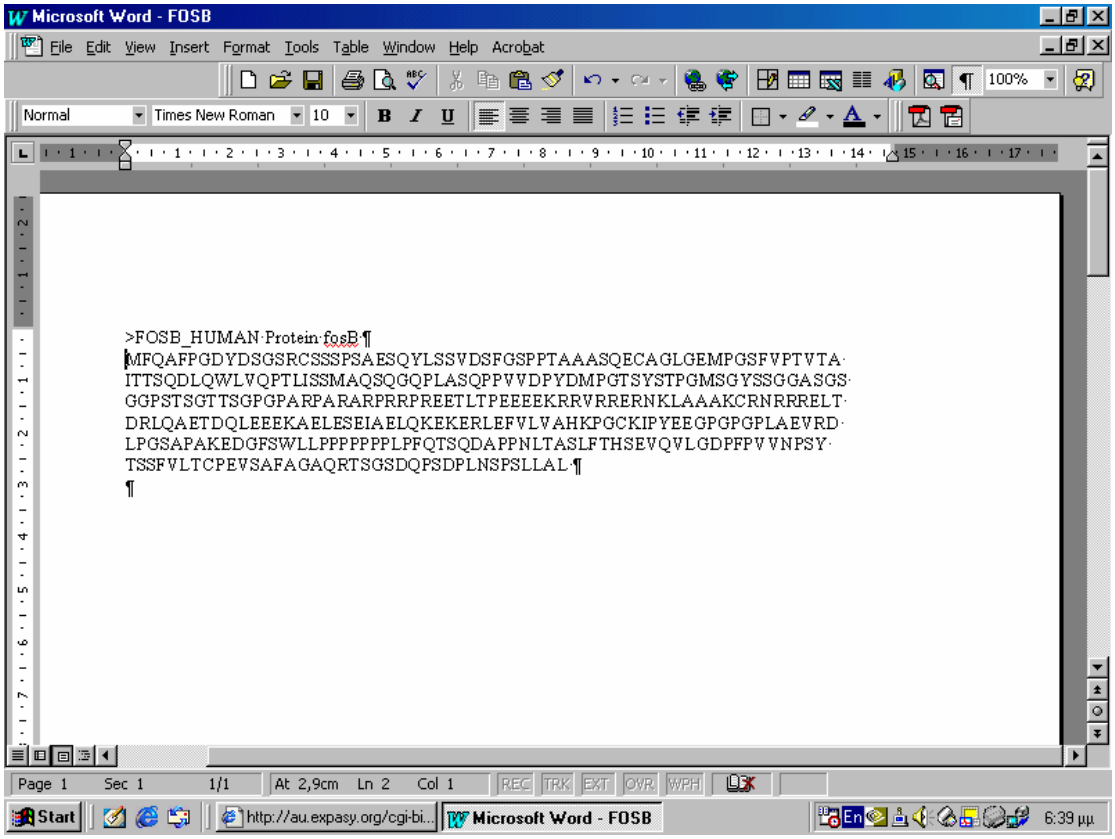
Επιλέξτε “FASTA format” για την ανάκτηση της αλληλουχίας.



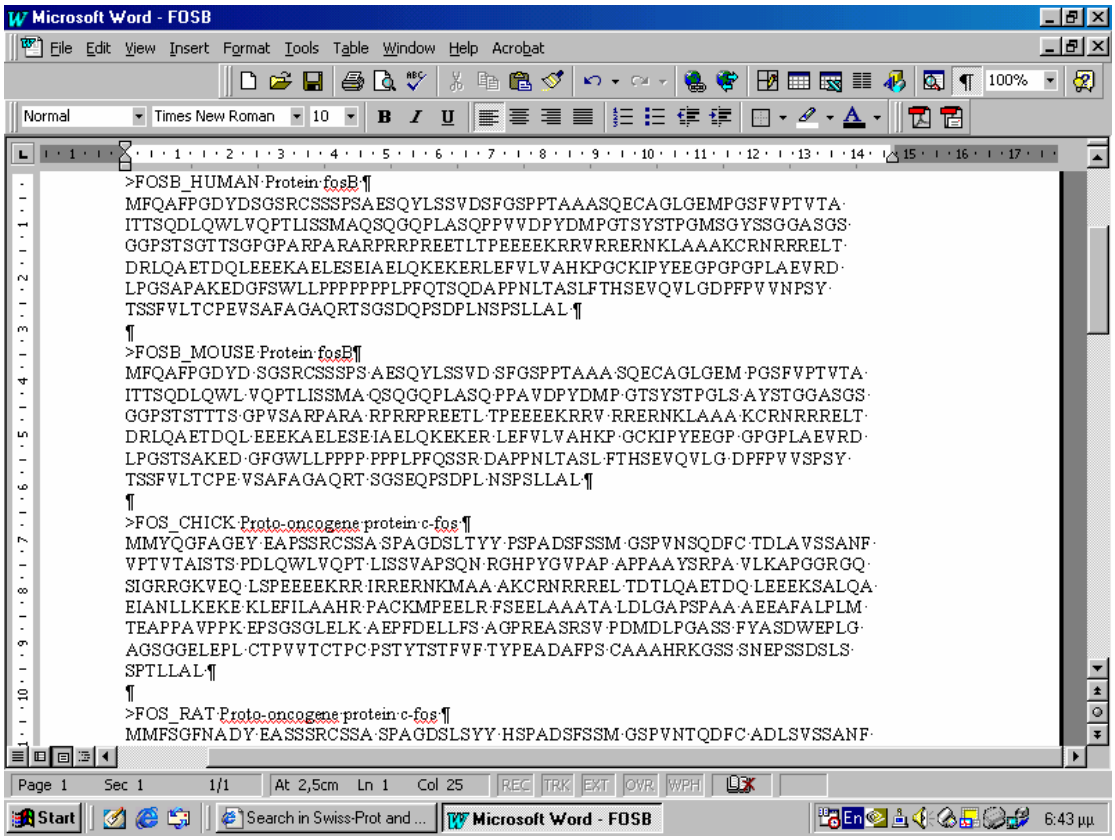
Αντιγράψτε τη αλληλουχία .



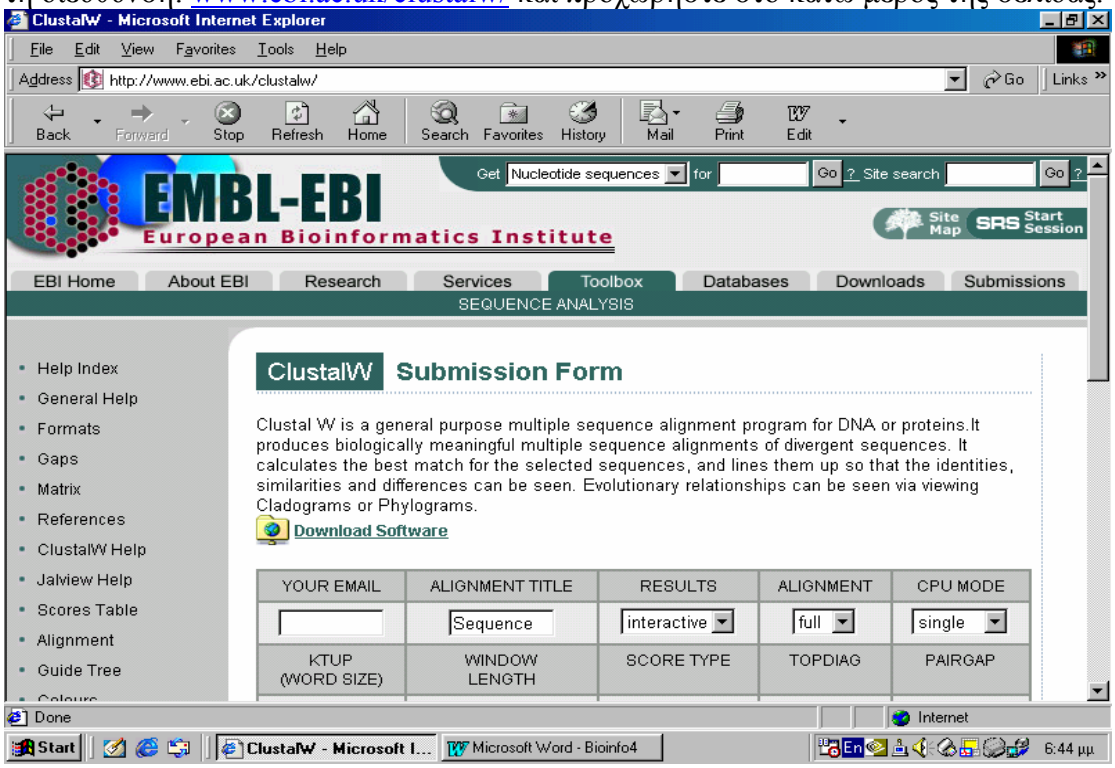
Ανοίξτε στο Microsoft Word ένα αρχείο, επικολλήστε την αντιγραμμένη αλληλουχία και σώστε το αρχείο ως "FOSB". Στην αρχή της αλληλουχίας προσθέστε την γραμμή >FOSB_HUMAN Protein fosB.



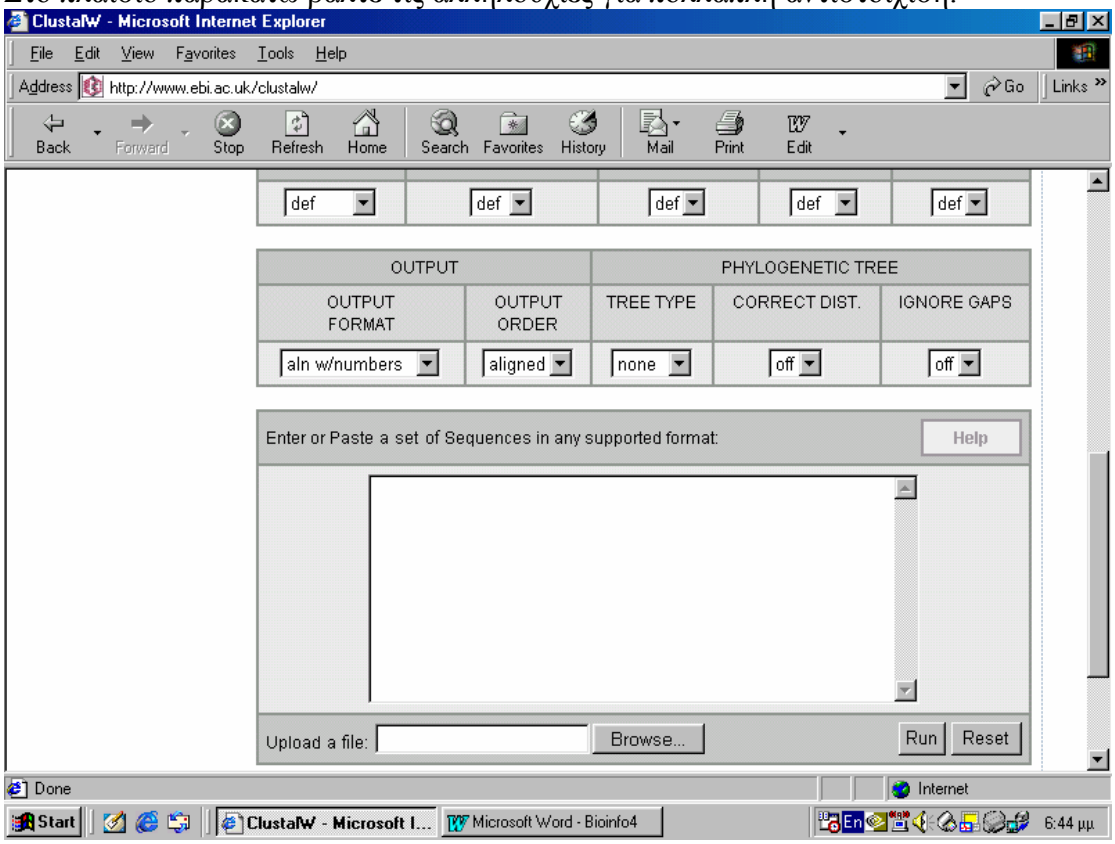
Επαναλάβετε τα ίδια βήματα και για τις αλληλουχίες FOSB_MOUSE, FOS_CHICK, FOS_RAT, FOS_MOUSE.



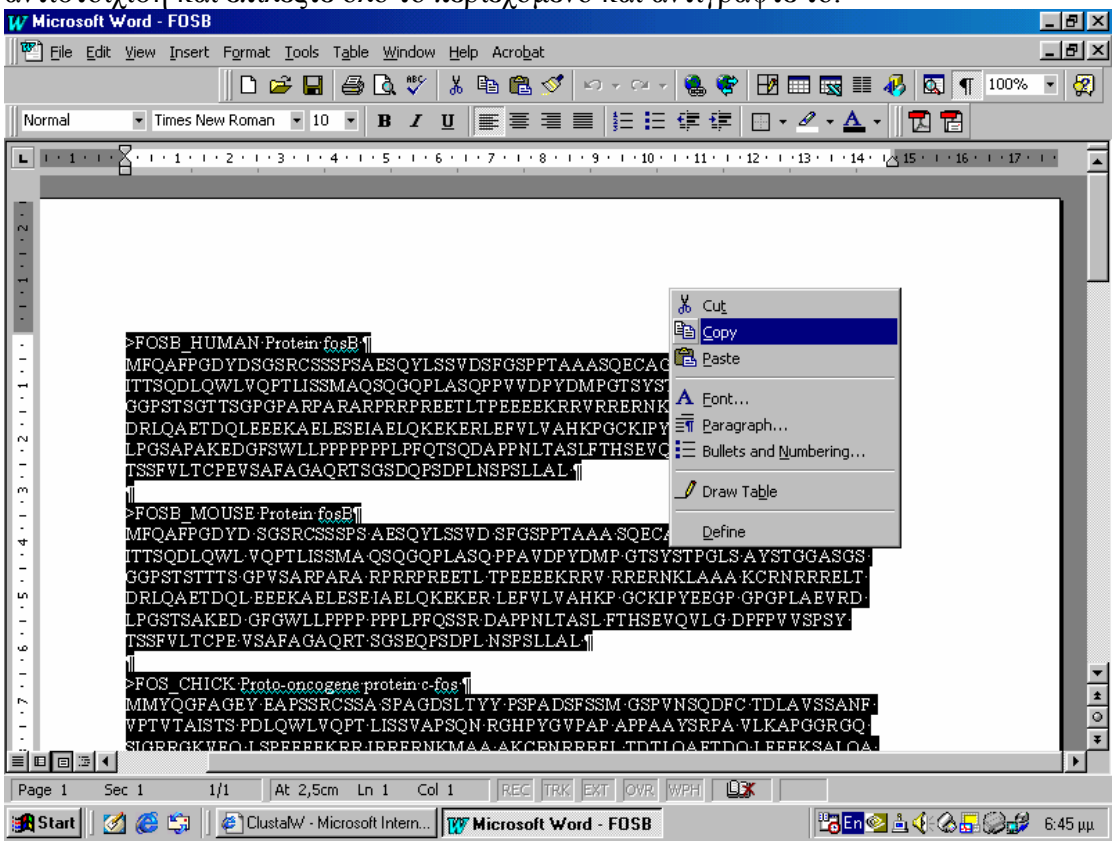
Για την πρόσβαση στο πρόγραμμα CrustalW, πληκτρολογήστε στο Internet Explorer τη διεύθυνση: www.ebi.ac.uk/clustalw/ και προχωρήστε στο κάτω μέρος της σελίδας.



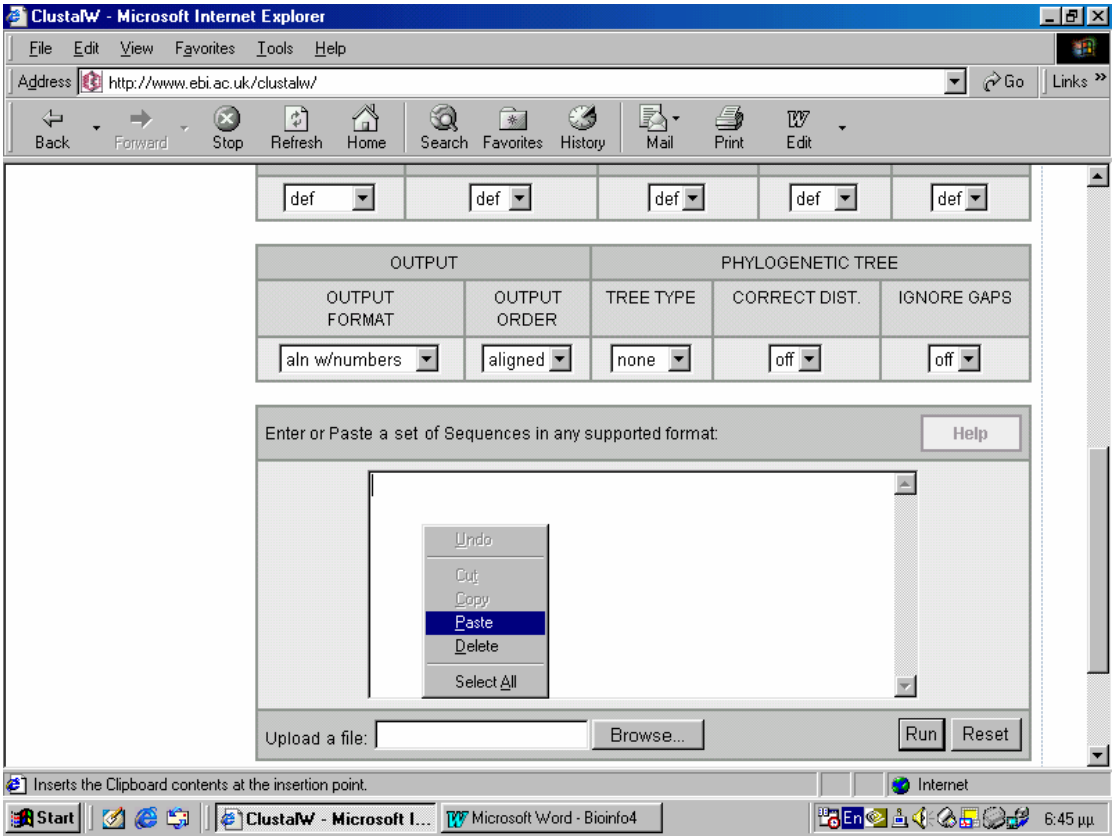
Στο πλαίσιο παρακάτω βάλτε τις αλληλουχίες για πολλαπλή αντιστοίχιση.



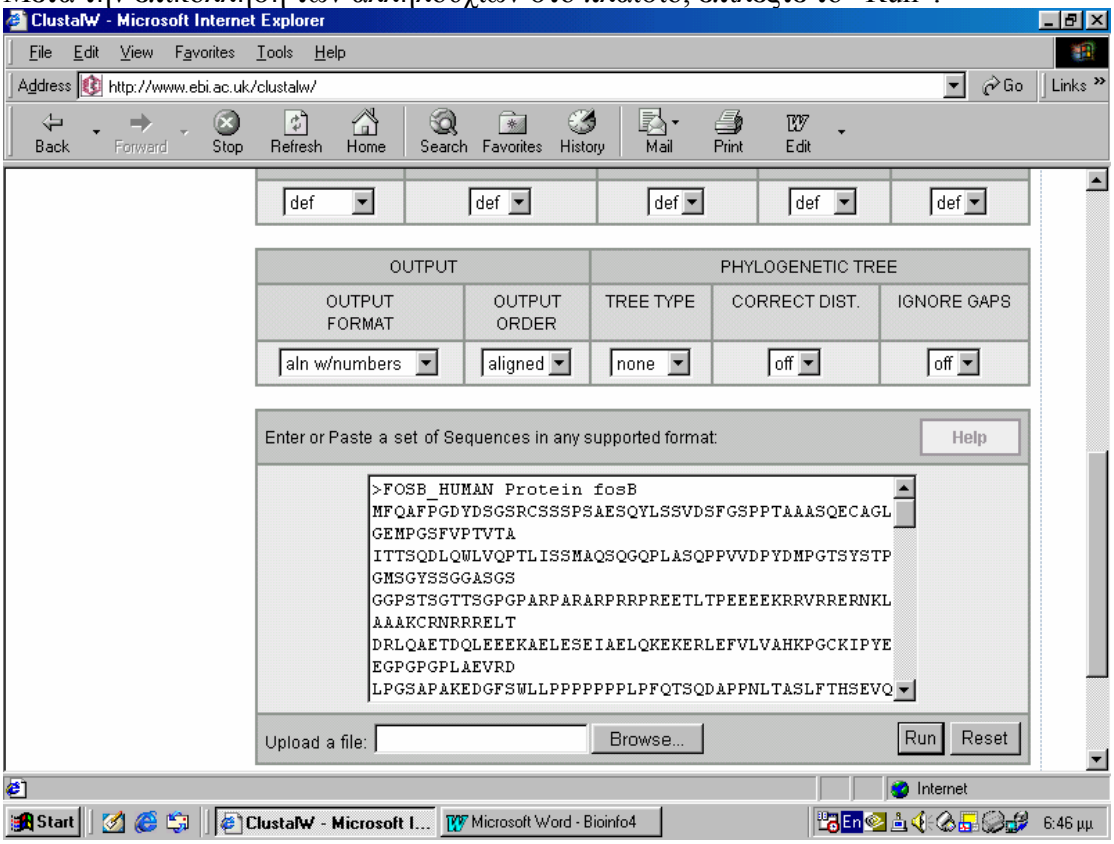
Ανοίξτε το αρχείο FOSB.doc στο Word, το οποίο περιέχει τις αλληλουχίες για αντιστοίχιση και επιλέξτε όλο το περιεχόμενο και αντιγράψτε το.



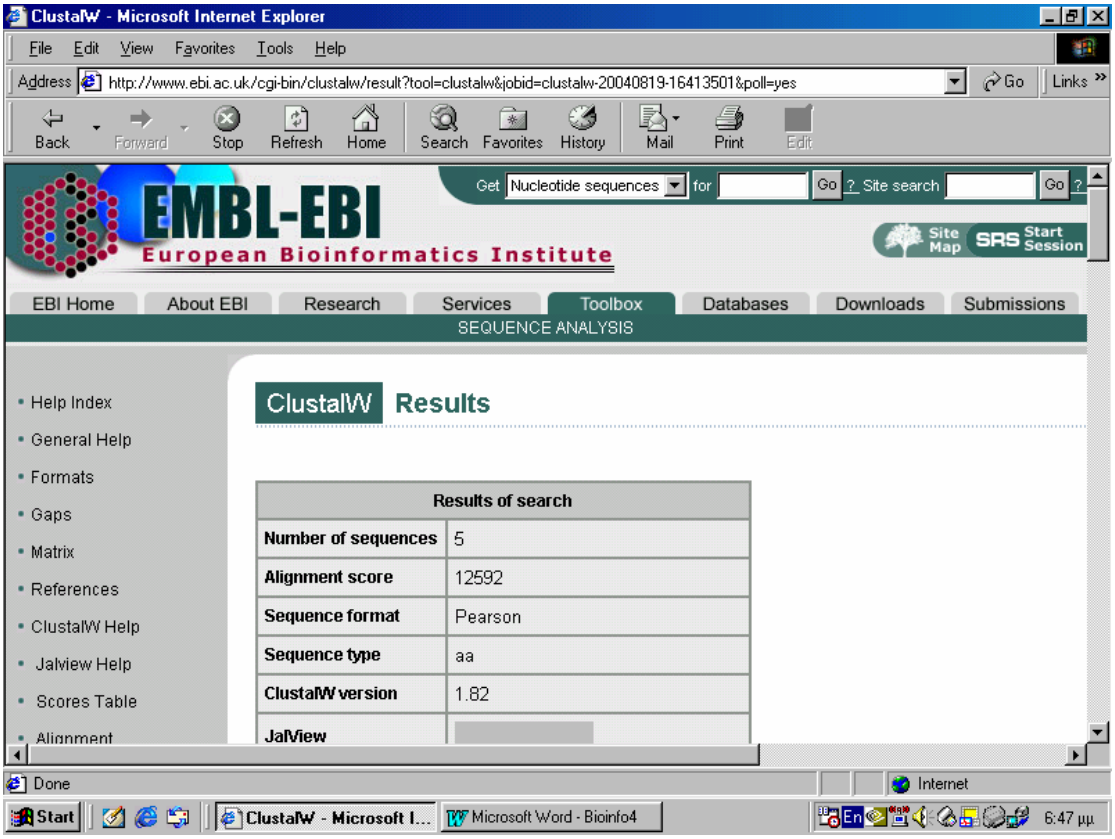
Στη συνέχεια, επικολλήστε τις αλληλουχίες από το αρχείο FOSB.doc στο παρακάτω πλαίσιο του προγράμματος ClustalW.



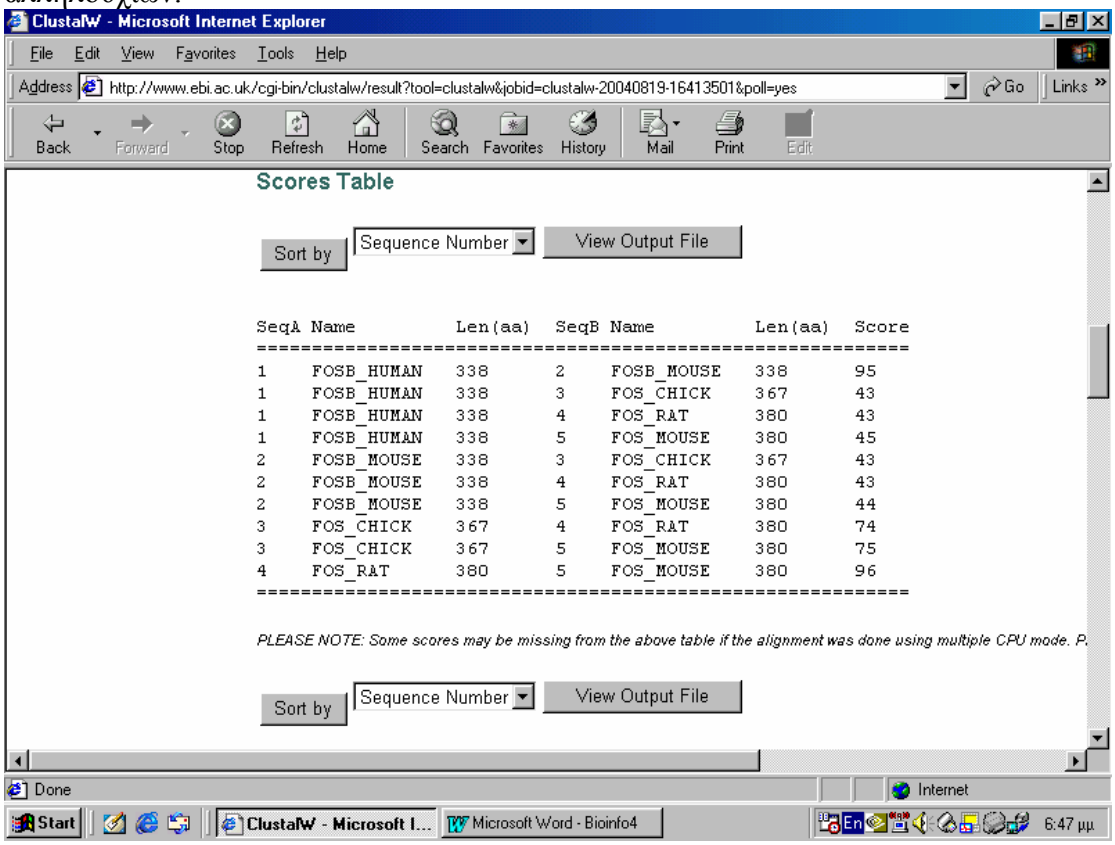
Μετά την επικόλληση των αλληλουχιών στο πλαίσιο, επιλέξτε το “Run”.



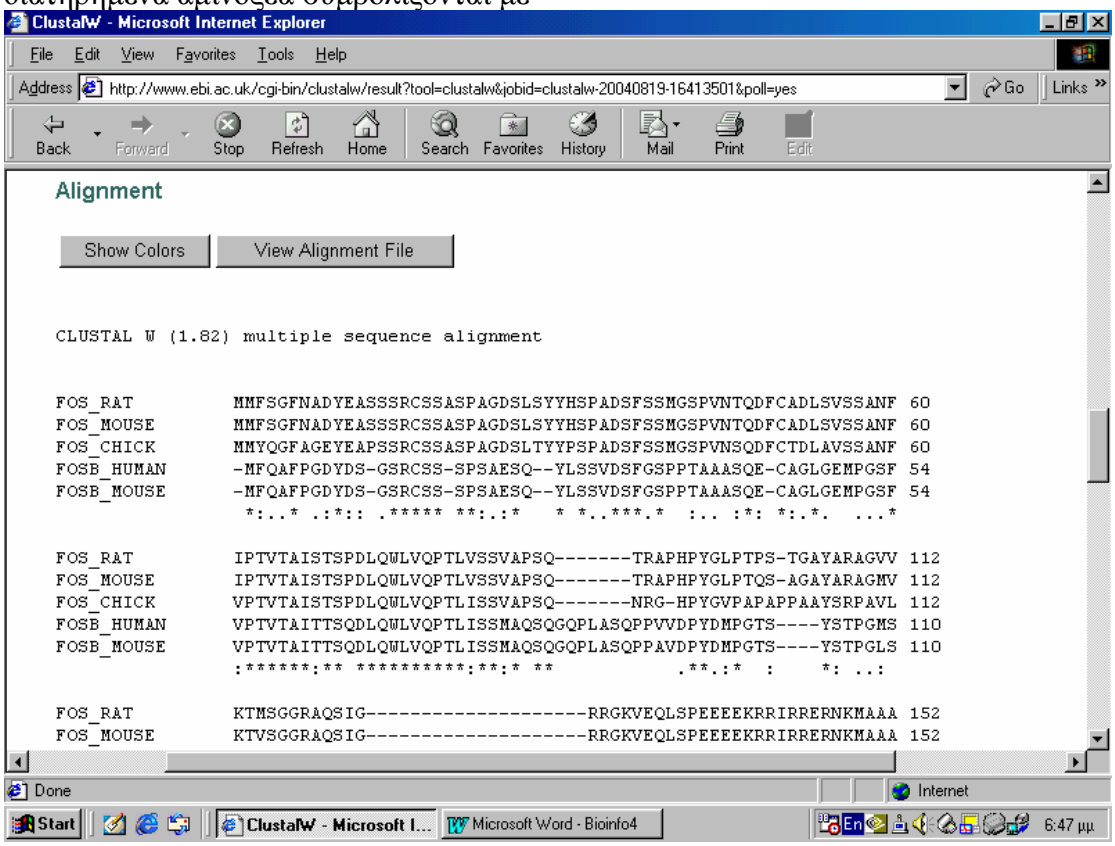
Τότε εμφανίζονται τα αποτελέσματα από την πολλαπλή αντιστοίχιση.

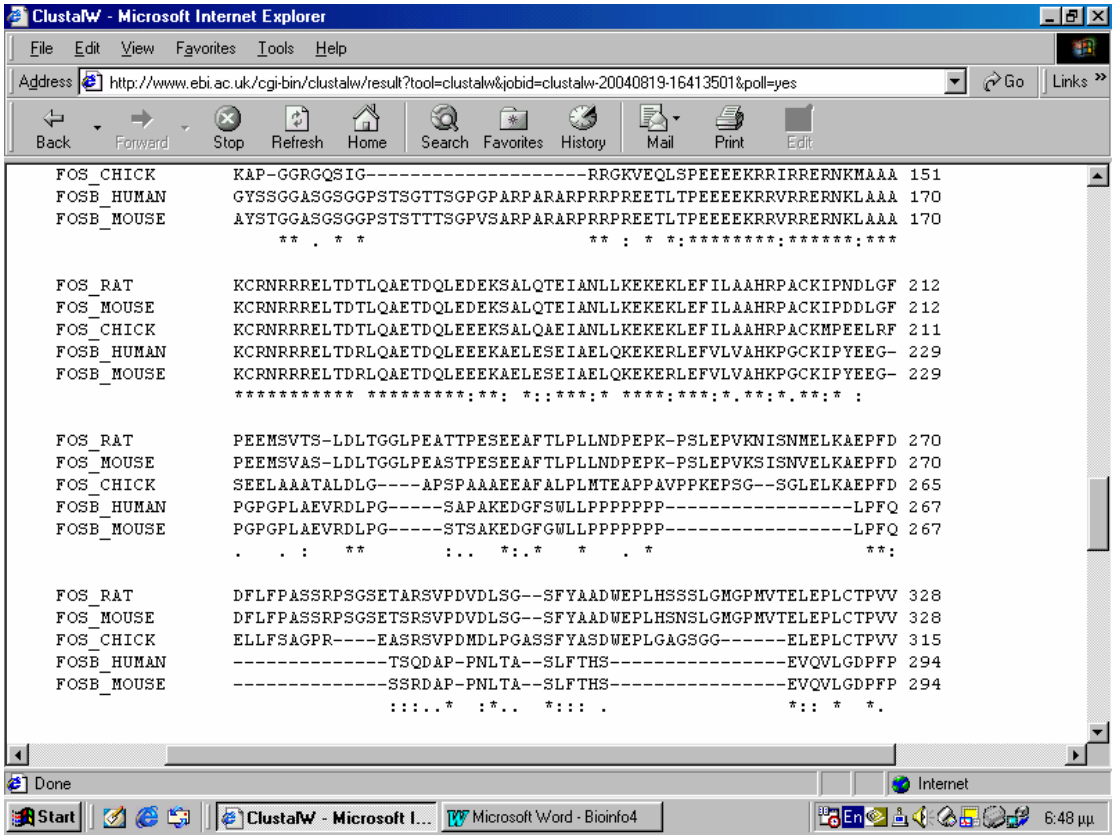


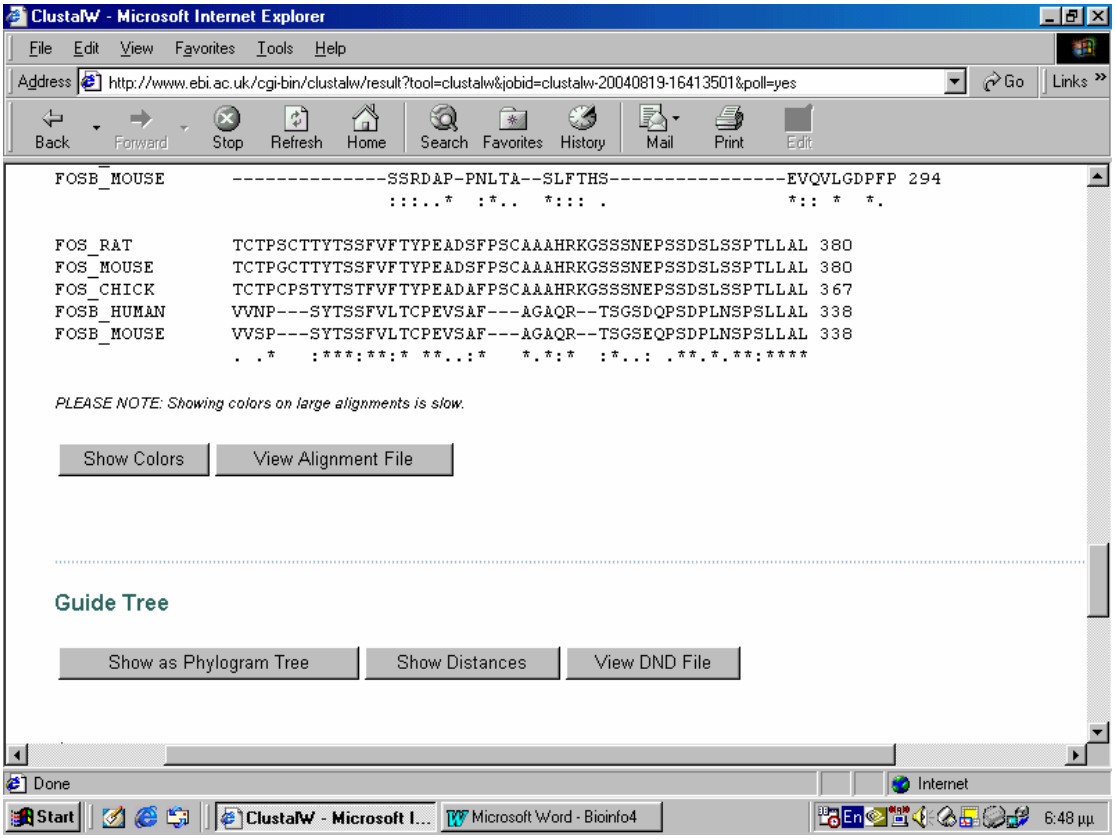
Επίσης, παρουσιάζονται οι βαθμολογίες ομοιοτήτων από τη σύγκριση ανά ζεύγη αλληλουχιών.



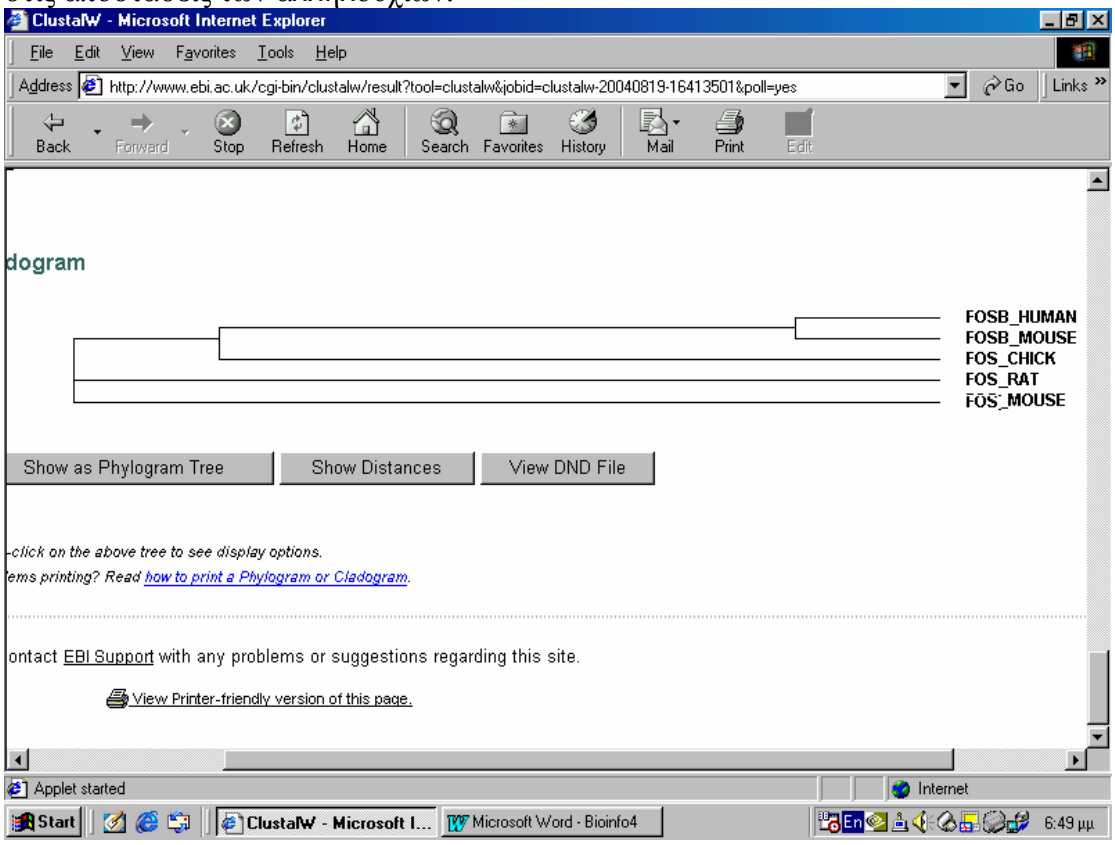
Επίσης, παρουσιάζονται τα αποτελέσματα από την πολλαπλή αντιστοίχιση. Τα διατηρημένα αμινοξέα συμβολίζονται με "*"



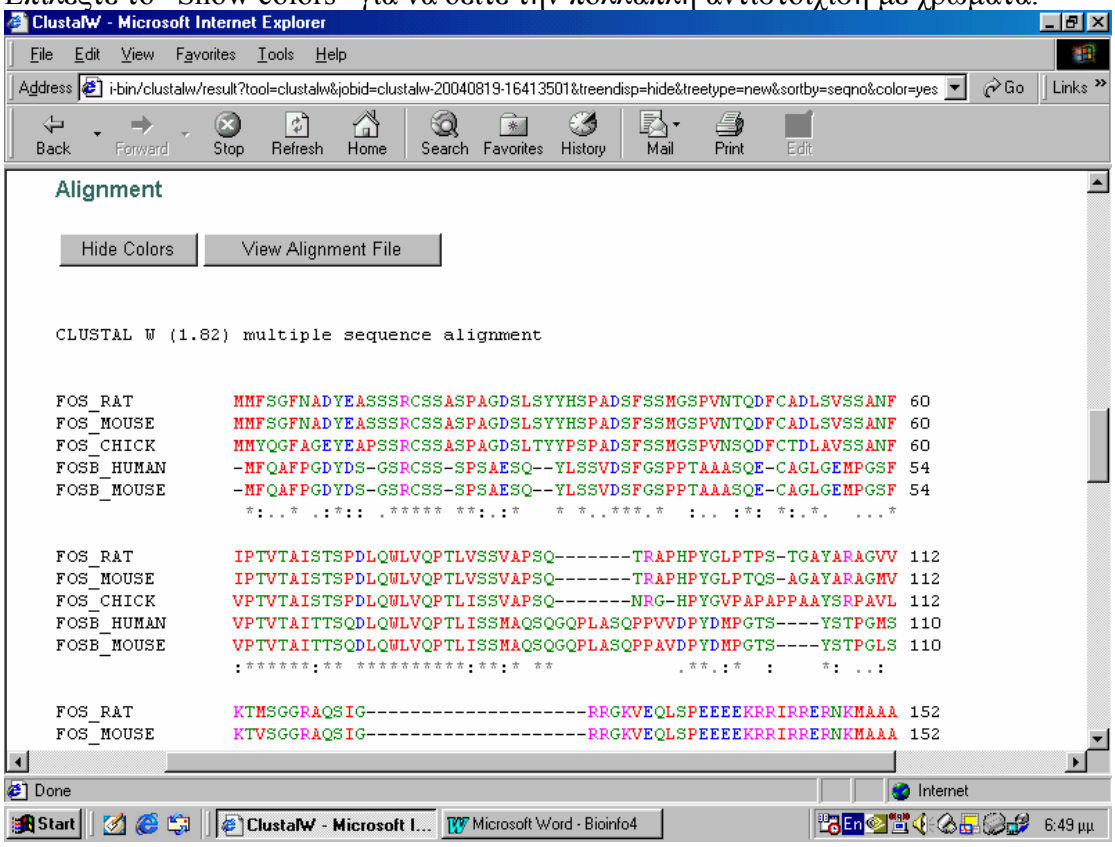

ClustalW - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.ebi.ac.uk/cgi-bin/clustalw/result?tool=clustalw&jobid=clustalw-20040819-16413501&poll=yes
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
FOS_CHICK KAP-GGRGQSIG----------RRGKVEQLSPEEEEKRRIRRERNKMAAA 151
FOSB_HUMAN GYSSGGASGSGGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAA 170
FOSB_MOUSE AYSTGGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAA 170
FOS_RAT KCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPNDLGF 212
FOS_MOUSE KCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGF 212
FOS_CHICK KCRNRRRELTDTLQAETDQLEEEKSALQAEIANLLKEKEKLEFILAAHRPACKMPEELRF 211
FOSB_HUMAN KCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEG- 229
FOSB_MOUSE KCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEG- 229
FOS_RAT PEEMSVTS-LDLTGGLPEATTPESEEAFTLPLLNDPEPK-PSLEPVKNISNMELKAEPFD 270
FOS_MOUSE PEEMSVAS-LDLTGGLPEASTPESEEAFTLPLLNDPEPK-PSLEPVKSISNVELKAEPFD 270
FOS_CHICK SEELAAATALDLG----APSPAAAEEAFALPLMTEAPPAVPPKEPSG--SGLELKAEPFD 265
FOSB_HUMAN PGPGPLAEVRDLPG-----SAPAKEDGFSWLLPPPPPPP----------LPFQ 267
FOSB_MOUSE PGPGPLAEVRDLPG-----STSAKEDGFGWLLPPPPPPP----------LPFQ 267
FOS_RAT DFLFPASSRPSGSETARSVPDVDLSG--SFYAADWEPLHSSSLGMGPMVTELEPLCTPVV 328
FOS_MOUSE DFLFPASSRPSGSETSRSVPDVDLSG--SFYAADWEPLHSNSLGMGPMVTELEPLCTPVV 328
FOS_CHICK ELLFSAGPR----EASRSVPDMDLPGASSFYASDWEPLGAGSGG------ELEPLCTPVV 315
FOSB_HUMAN ----------TSQDAP-PNLTA--SLFTHS----------EVQVLGDPFP 294
FOSB_MOUSE ----------SSRDAP-PNLTA--SLFTHS----------EVQVLGDPFP 294
Done
Internet
Start ClustalW - Microsoft I... Microsoft Word - Bioinfo4 6:48 μμ


ClustalW - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.ebi.ac.uk/cgi-bin/clustalw/result?tool=clustalw&jobid=clustalw-20040819-16413501&poll=yes
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
FOSB_MOUSE ----------SSRDAP-PNLTA--SLFTHS----------EVQVLGDPFP 294
FOS_RAT TCTPSCTTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
FOS_MOUSE TCTPGCTTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
FOS_CHICK TCTPCPSTYTSTFVFTYPEADAFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 367
FOSB_HUMAN VVNP---SYTSSFVLTCPEVSAF---AGAQR--TSGSDQPSDPLNSPSLLAL 338
FOSB_MOUSE VVSP---SYTSSFVLTCPEVSAF---AGAQR--TSGSEQPSDPLNSPSLLAL 338
PLEASE NOTE: Showing colors on large alignments is slow.
Show Colors
View Alignment File
Guide Tree
Show as Phylogram Tree
Show Distances
View DND File
Done
Internet
Start ClustalW - Microsoft I... Microsoft Word - Bioinfo4 6:48 μμ

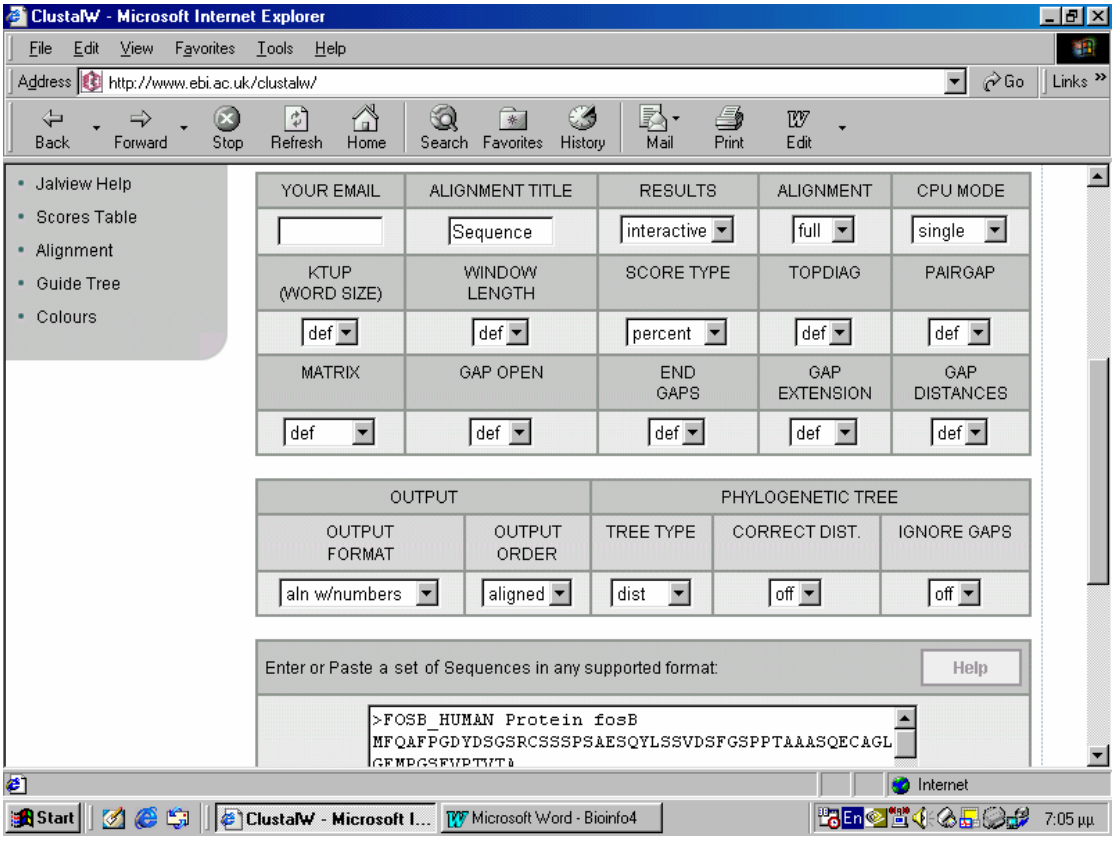
Κυλήστε την οθόνη προς τα κάτω για να δείτε φυλογενετικό δένδρο, που αντιστοιχεί στις αποστάσεις των αλληλουχιών.

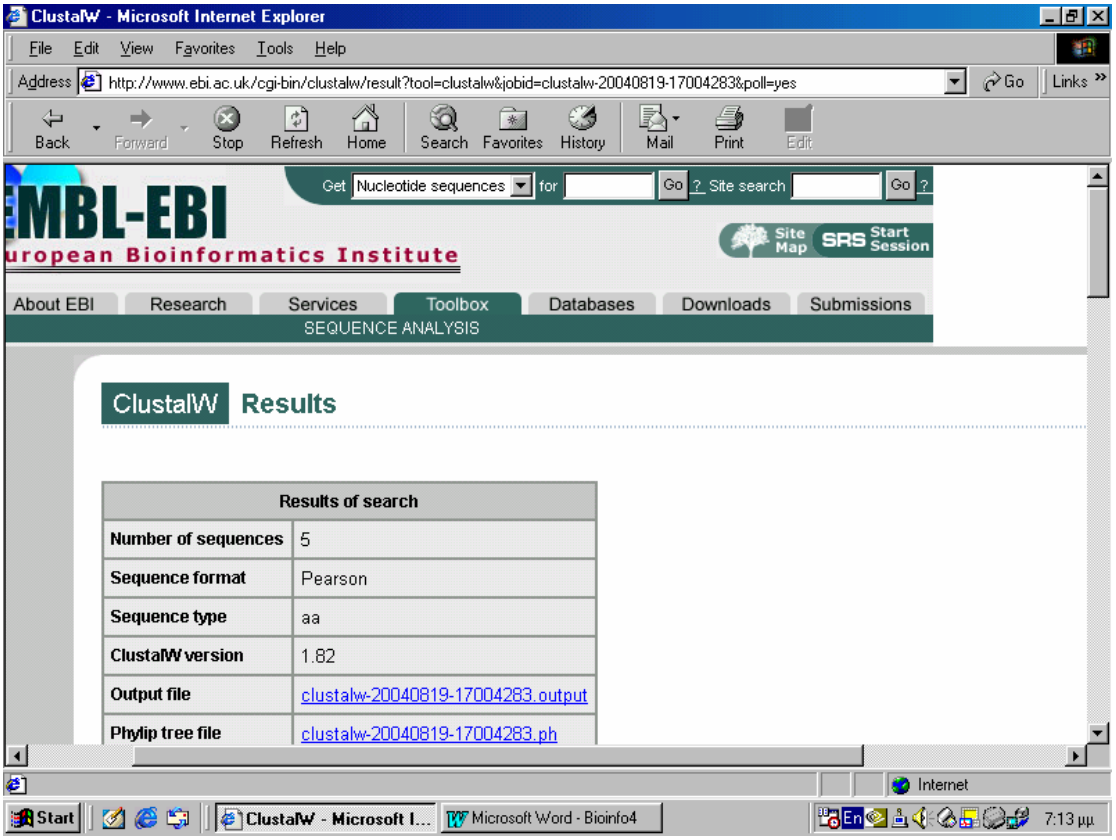


Επιλέξτε το "Show colors" για να δείτε την πολλαπλή αντιστοίχιση με χρώματα.



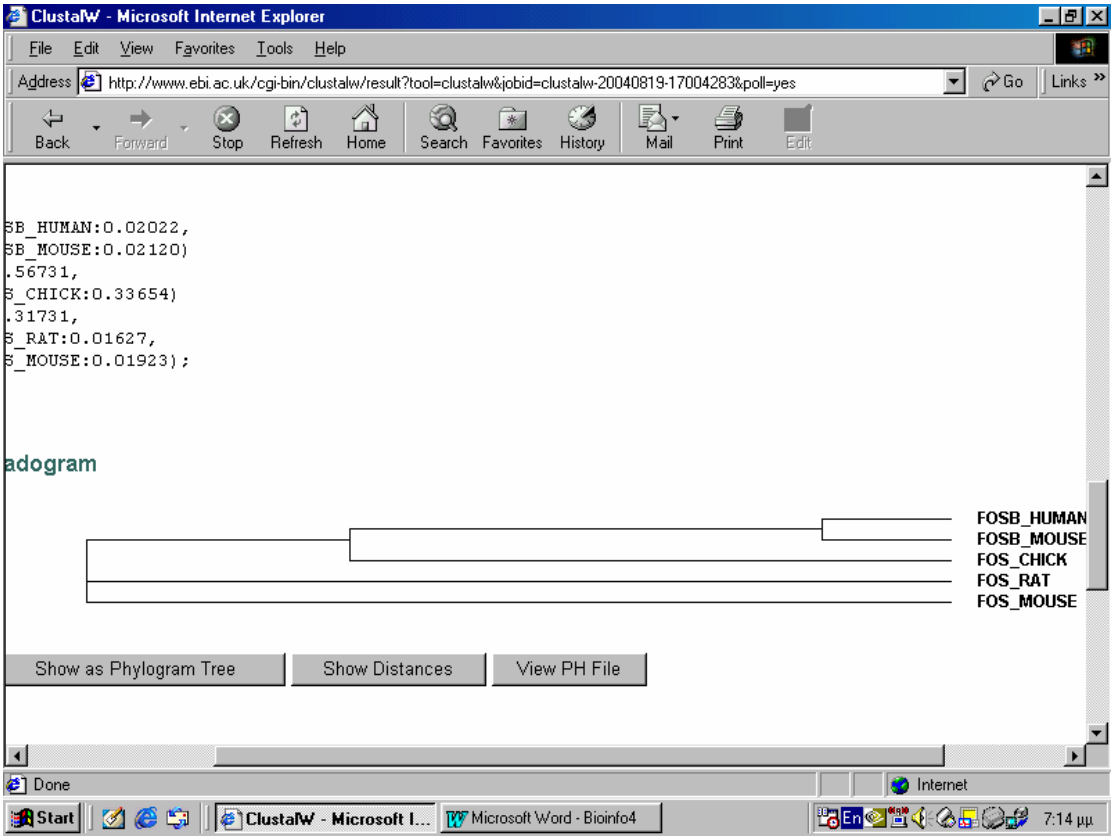
Κάντε κλικ στο “Back” για να επιστρέψτε στην αρχική σελίδα του ClustalW και στο



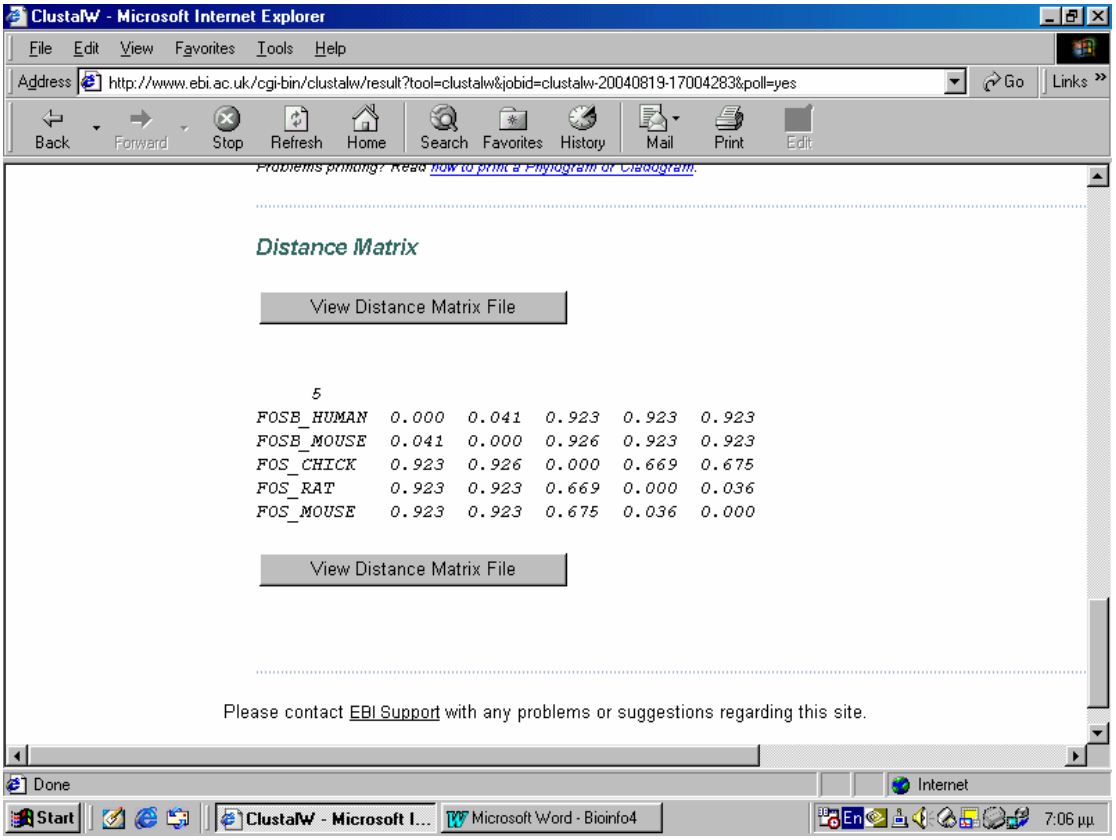


“TREE TYPE” επιλέξτε “dist” και επιλέξτε το “Run”.

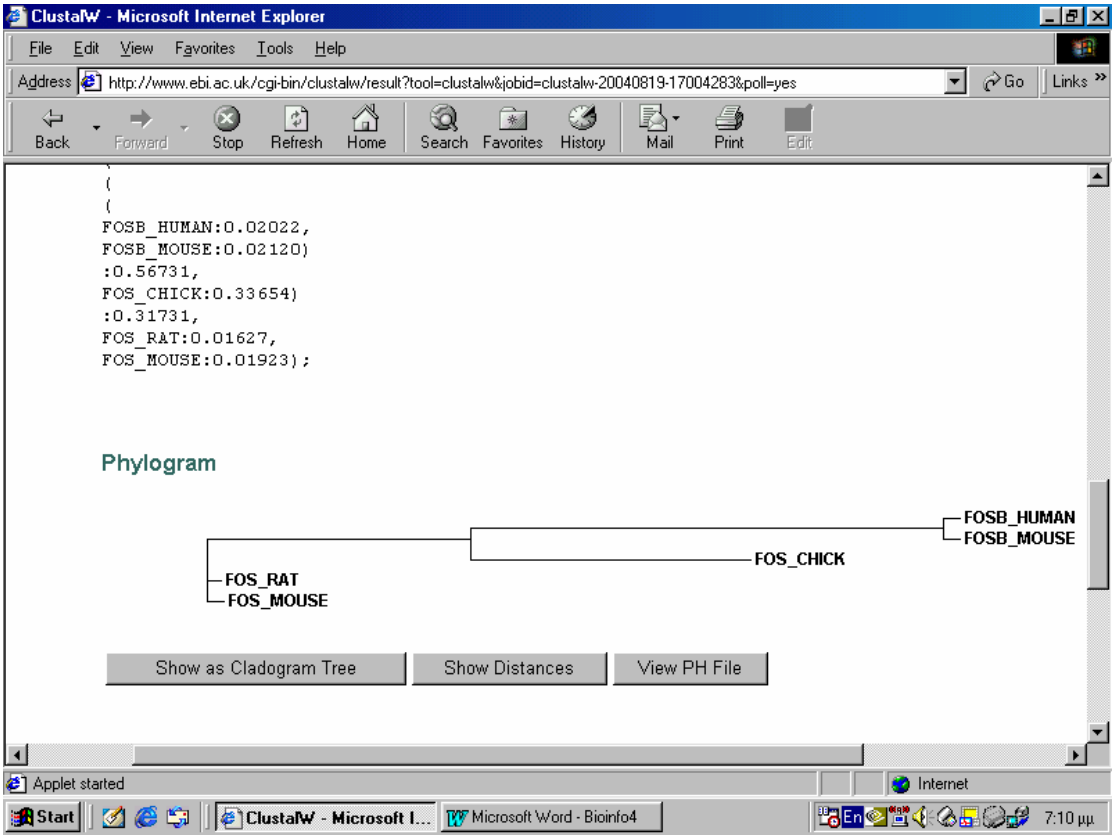
Μετά προχωρήστε προς τα κάτω για να δείτε τις αποστάσεις μεταξύ των αλληλουχιών.



Παρουσιάζονται οι αποστάσεις για την σύνθεση του καθοδηγητικού δέντρου που χρειάζεται για την πολλαπλή αντιστοίχιση.

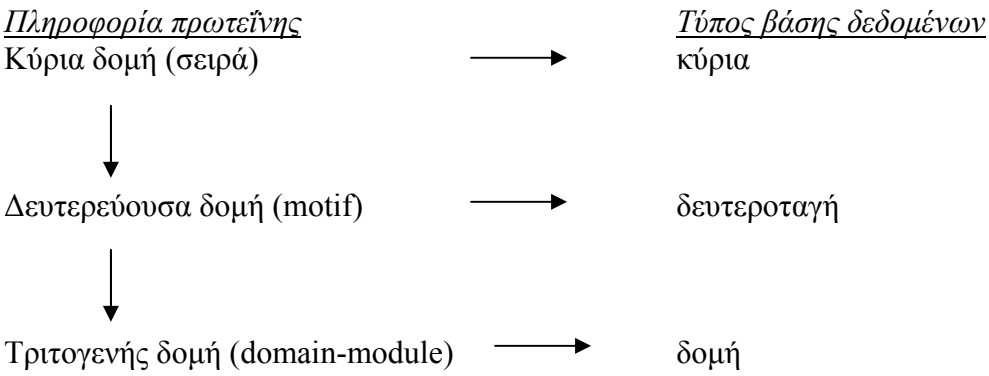


Ο τύπος του δέντρου υποδηλώνει την εξελικτική σχέση μεταξύ των αλληλουχιών.



Πηγές δεδομένων για πρωτεΐνες

Υπάρχει ένα εύρος πηγών πληροφοριών για τις πρωτεΐνες, οι οποίες χωρίζονται ανάλογα με την πληροφορία που παρέχουν, δηλ. τον τύπο δεδομένων των πρωτεϊνών: βάσεις δεδομένων με κύρια στοιχεία (π.χ. αλληλουχίες), βάσεις δεδομένων με δευτεροταγή δεδομένα και βάσεις δεδομένων με δομές πρωτεϊνών.
Ο τύπος της βάσης δεδομένων ορίζεται ανάλογα με το επίπεδο της πληροφορίας για τις πρωτεΐνες που έχουν αποθηκευμένο:



Οι κύριες βάσεις δεδομένων περιέχουν την κύρια δομή της πρωτεΐνης, δηλ. τη σειρά αμινοξέων που συμβολίζονται με γράμματα στη σειρά.
Στις δευτεροταγής βάσεις δεδομένων υπάρχουν (patterns) (συνηθισμένες εκφράσεις, «δαχτυλικά αποτυπώματα», μπλόκα, περιγράμματα), δηλ. οι δευτεροταγές δομές αντιστοιχούν σε περιοχές με τοπική κανονικότητα (a-helices και b-strands) που είναι διατηρημένα μοτίβα.
Στις βάσεις δεδομένων δομής υπάρχουν τομείς (modules) που είναι τα αποτελέσματα της συσκευασίας των δευτερεύουσων δομών μέσα σε μια δίπλωση (αυτόνομες διπλούμενες μονάδες). Η πληροφορία υπάρχει ως ένα σύνολο ατομικών συντεταγμένων.

Κύριες βάσεις δεδομένων για αλληλουχίες

Στην επιστημονική βιβλιογραφία, όταν οι πληροφορίες για αλληλουχίες άρχισαν να γίνονται πολλές, τότε ξεκίνησαν σε διάφορα μέρη του κόσμου πολλές μελέτες για τη εξέλιξη κυρίων βάσεων δεδομένων. Οι κύριες βάσεις δεδομένων με αλληλουχίες πρωτεϊνών είναι οι παρακάτω:

PIR
MIPS
SWISS-PROT
TrEMBL
NRL-3D

PIR

To Potein Information Resource (PIR) δημιουργήθηκε στο National Biomedical Research Foundation (NBRF) από την M. Dayhoff, σαν μια συλλογή αλληλουχιών για την διερεύνηση των εξελικτικών σχέσεων ανάμεσα στις πρωτεΐνες.

Στην σημερινή της μορφή, η βάση δεδομένων διαχωρίζεται σε 4 τμήματα: PIR1-PIR4, που διαφέρουν όσον αφορά την ποιότητα των δεδομένων και το επίπεδο σχολιασμού που δίνεται. Το PIR1 περιέχει πλήρεις ταξινομημένες καταγραφές με σχόλια και το PIR4 περιέχει καταχωρήσεις που είναι εννοιολογικές μεταταφράσεις των αλληλουχιών.

MIPS

Το Martinsried Institute for Protein Sequences (MIPS) επιλέγει και επεξεργάζεται δεδομένα αλληλουχιών για την βάση δεδομένων του PIR.

SWISS-PROT

Η βάση δεδομένων SWISS-PROT δημιουργήθηκε από το τμήμα Ιατρικής Βιοχημείας στο Πανεπιστήμιο της Γενεύης και από το EMBL, τώρα η βάση δεδομένων διατηρείται σε συνεργασία με το Ελβετικό Ινστιτούτο Βιοπληροφορικής (SIB) και το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI) του EMBL.
Η βάση δεδομένων δίνει πολλά σχόλια, συμπεριλαμβάνοντας περιγραφές λειτουργιών της πρωτεΐνης και της δομής των τομέων της, των παραλλαγών της κ.ο.κ. Είναι ελάχιστα πλεονάζουσα και είναι διασυνδεόμενη με πολλές άλλες πηγές πληροφόρησης.

TrEMBL

Το 1996, η TrEMBL, δημιουργήθηκε ως συμπλήρωμα για τη SWISS-PROT. Περιέχει τις μεταφράσεις όλων των κωδικοποιημένων αλληλουχιών στο EMBL. Οι καταχωρήσεις εμπεριέχονται στη SWISS-PROT αλλά με ανεπαρκείς αναλύσεις και σχόλια.

NRL-3D

Η βάση δεδομένων NRL-3D δημιουργήθηκε από το PIR, από αλληλουχίες που ανακτήθηκαν από τη Brookhaven Protein Bank (PBD). Στη PBD, οι πληροφορίες για αλληλουχίες είναι διαθέσιμες για ανάκτηση με λέξεις κλειδιά και για αναζήτηση ομοιοτήτων μεταξύ αλληλουχιών.

Η δομή των καταχωρήσεων του SWISS-PROT
Η SWISS-PROT είναι η πιο δημοφιλής βάση δεδομένων για θέματα αναζήτησης λόγω της δομής της και της ποιότητας των σχολίων της. Περιέχει περίπου 100000 καταχωρήσεις από πάνω από περίπου 5000 διαφορετικά είδη.

Μια καταχώρηση της SWISS-PROT έχει την παρακάτω μορφή:

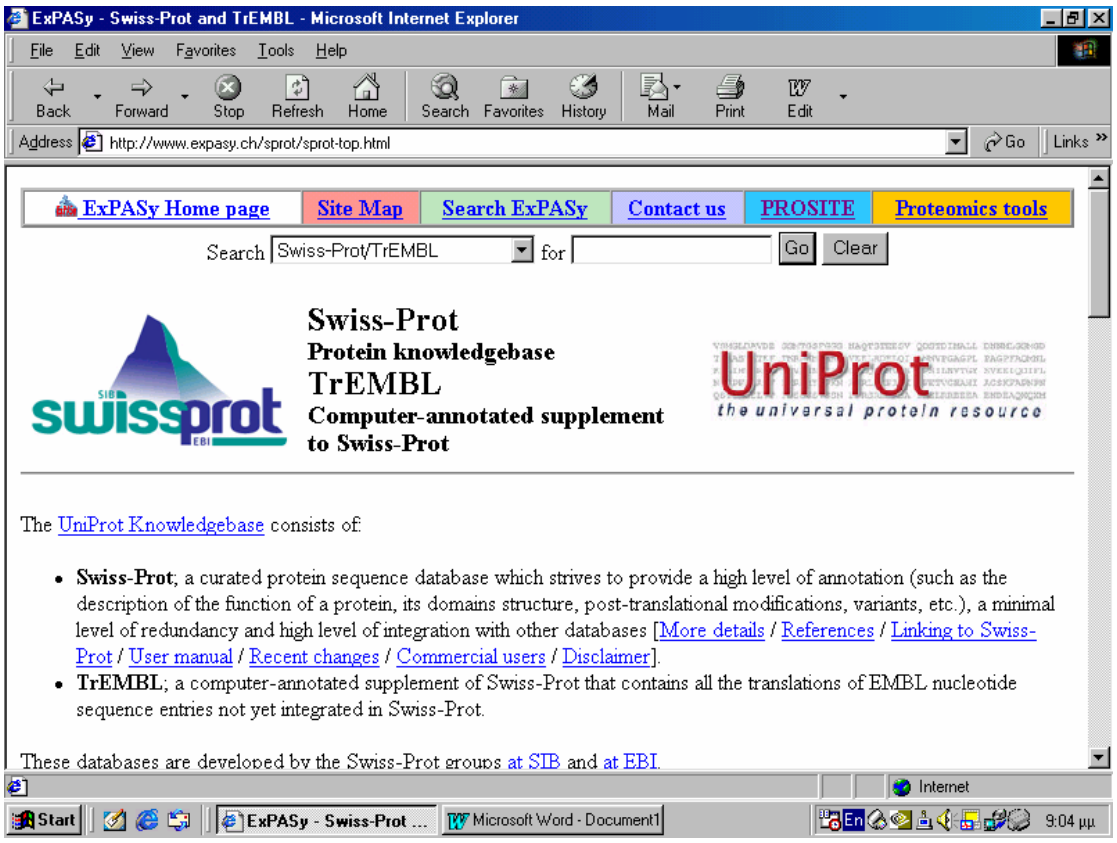
```
ID   OPSD_SHEEP     STANDARD;      PRT;   348 AA.
AC   P02700;
DT   21-JUL-1986 (Rel. 01, Created)
DT   01-FEB-1991 (Rel. 17, Last sequence update)
DT   05-JUL-2004 (Rel. 44, Last annotation update)
DE   Rhodopsin.
GN   Name=RHO;
OS   Ovis aries (Sheep).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae;
OC   Caprinae; Ovis.
OX   NCBI_TaxID=9940;
RN   [1]
RP   SEQUENCE.
RA   Pappin D.J.C., Elipoulos E., Brett M., Findlay J.B.C.;
RT   "A structural model for ovine rhodopsin.";
RL   Int. J. Biol. Macromol. 6:73-76(1984).
...
CC   -!- FUNCTION: Visual pigments are the light-absorbing molecules that
CC       mediate vision. They consist of an apoprotein, opsin, covalently
CC       linked to cis-retinal.
CC   -!- SUBCELLULAR LOCATION: Integral membrane protein.
CC   -!- TISSUE SPECIFICITY: Rod shaped photoreceptor cells which mediates
CC       vision in dim light.
CC   -!- MISCELLANEOUS: Maximal absorption at 495 nm.
CC   -!- SIMILARITY: Belongs to family 1 of G-protein coupled receptors.
CC       Opsin subfamily.
DR   PIR; ; OOSH.
...
DR   Pfam; PF00001; 7tm_1; 1.
DR   PRINTS; PR00237; GPCRRHODOPSN.
DR   PRINTS; PR00238; OPSIN.
DR   PRINTS; PR00579; RHODOPSIN.
DR   PROSITE; PS00237; G_PROTEIN_RECEP_F1_1; 1.
DR   PROSITE; PS50262; G_PROTEIN_RECEP_F1_2; 1.
DR   PROSITE; PS00238; OPSIN; 1.
KW   Direct protein sequencing; G-protein coupled receptor; Glycoprotein;
KW   Lipoprotein; Palmitate; Phosphorylation; Photoreceptor;
KW   Retinal protein; Transmembrane; Vision.
FT   DOMAIN        1     36       Extracellular.
FT   TRANSMEM     37     61       1 (Potential).
FT   DOMAIN       62     73       Cytoplasmic.
FT   TRANSMEM     74     98       2 (Potential).
FT   DOMAIN       99    113       Extracellular.
...
FT   BINDING     296    296       Retinal chromophore.
FT   LIPID       322    322       S-palmitoyl cysteine (By similarity).
FT   LIPID       323    323       S-palmitoyl cysteine (By similarity).
FT   DISULFID    110    187       By similarity.
FT   MOD_RES     334    334       Phosphoserine (by RK).
FT   MOD_RES     335    335       Phosphothreonine (by RK).
...
SQ   SEQUENCE   348 AA;  38891 MW;  AAFD6F0D6A8BAEE5 CRC64;
     MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIVL GFPINFLTLY
     VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH GYFVFGPTGC NLEGFFATLG
     GEIALWSLVV LAIERYVVVC KPMSNFRFGE NHAIMGVAFT WVMALACAAP PLVGWSRYIP
     QGMQCSCGAL YFTLKPEINN ESFVIYMFVV HFSIPLIVIF FCYGQLVFTV KEAAAQQQES
     ATTQKAEKEV TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKSSSV
     YNPVIYIMMN KQFRNCMLTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
//
```

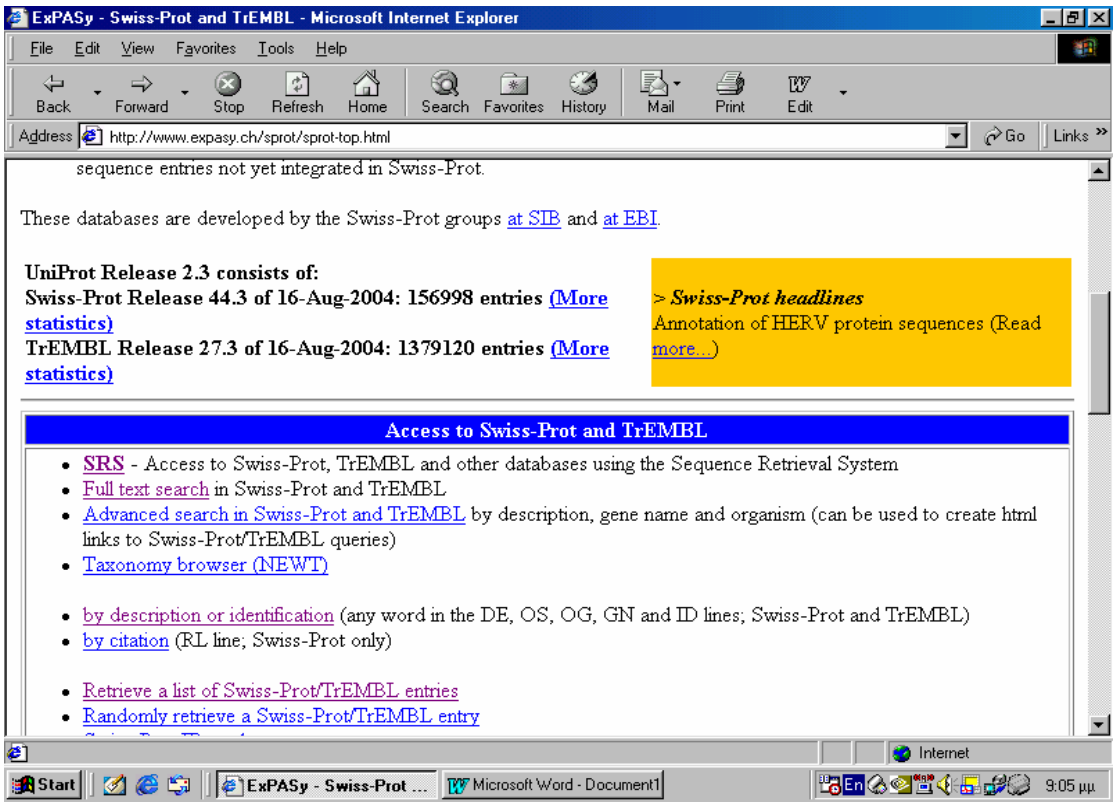
Μια καταχώρηση του SWISS-PROT έχει την παρακάτω μορφή:
Κάθε γραμμή είναι σημειωμένη με ένα διψήφιοκωδικό, που βοηθάει στην παρουσίαση της πληροφορίας με ένα δομημένο τρόπο.
Η γραμμή **ID** μας πληροφορεί ότι το όνομα της καταχώρησης είναι OPSD_SHEEP, μια πρωτεΐνη με 348 αμινοξέα. Ο τύπος του ID είναι PROTEIN_SOURCE, δηλ. ο τύπος της πρωτεΐνης και η ονομασία του οργανισμού.
Το **AC** (P02700) είναι ο κωδικός πρόσβασης που συνήθως παραμένει σταθερός μεταξύ των βάσεων δεδομένων. Το AC είναι απαραίτητο διότι μερικές φορές το ID αλλάζει.
Τα πεδία **DT** δίνουν πληροφορίες πάνω στην ημερομηνία καταχώρησης της σειράς στη βάση δεδομένων, όπως και πότε τροποποιήθηκε.
Τα πεδία **DE** μας πληροφορούν τις ονομασίες με τις οποίες οι πρωτεΐνες είναι γνωστές, π.χ. rhodopsin.
Το πεδίο **GN** δίνει την ονομασία του γονίδιου
Το πεδίο **OS** δίνει το είδος του οργανισμού
Το πεδίο **OC** δίνει την ταξινόμηση του οργανισμού μέσα στα βιολογικά βασίλεια.
Τα πεδία **RN, RP, RA, RL** δίνουν αναφορές
Τα πεδία **CC** παρέχουν σχόλια για τη λειτουργία της πρωτεΐνης, τις μετά-μεταφραστικές αλλαγές (PTM), τον τύπο του ιστού, την υπό-κυτταρική τοποθεσία και την ομοιότητα ή τον δεσμό με συγκεκριμένες οικογένειες πρωτεϊνών.
Τα πεδία **DR** παρέχουν συνδέσμους με άλλες βιομοριακές βάσεις δεδομένων
Τα πεδία **KW** παρέχουν λέξεις κλειδιά
Τα πεδία FT σχηματίζουν ένα πίνακα χαρακτηριστικών (Feature Table), ο οποίος επισημαίνει περιοχές ενδιαφέροντος στην αλληλουχία: τοπικά δευτεροταγή δομή (όπως είναι οι transmembrane τομείς), θέσεις στις οποίες ο συνθέτης συνδέεται με το υπόστρωμα (ligand binding sites), κλπ. Κάθε πεδίο περιέχει μια λέξη κλειδί (π.χ. TRANSMEM), την τοποθεσία του χαρακτηριστικού στην αλληλουχία (π.χ. 37-61), και ένα σχόλιο που μπορεί να υποδηλώνει το επίπεδο αξιοπιστίας ενός συγκεκριμένου σχολιασμού (π.χ. POTENTIAL). Οι εκχωρήσεις των transmembrane τομέων προκύπτουν από την εφαρμογή λογισμικού πρόβλεψης και έτσι με την έλλειψη τρισδιάστατων πειραματικών δεδομένων δομής, μπορούν να προσδιοριστούν μόνο σαν πιθανοί (potential).
Τα πεδία **SQ** περιέχουν την καθεαυτό σειρά. Ένας κωδικός που αποτελείται από ένα γράμμα του αμινοξέου και κάθε γραμμή περιέχει 60 αμινοξέα. Το **MW** δίνει το μοριακό βάρος.

ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

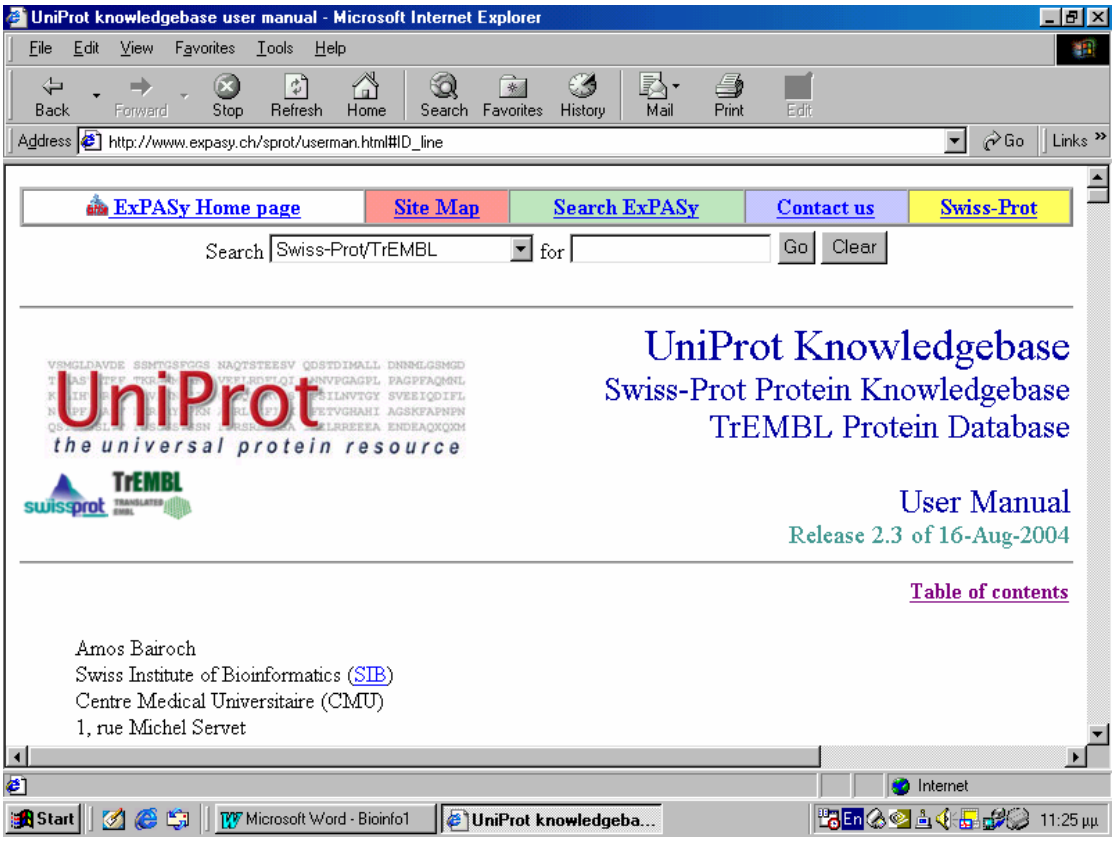
Καταχώρηση από το SWISS-PROT

Ανοίξτε το Internet Explorer για πρόσβαση στο SWISS-PROT, γράφοντας τη διεύθυνση www.expasy.ch/sprot/sprot-top.htm. Οπότε θα εμφανισθεί η παρακάτω σελίδα, κυλήστε την προς τα κάτω.



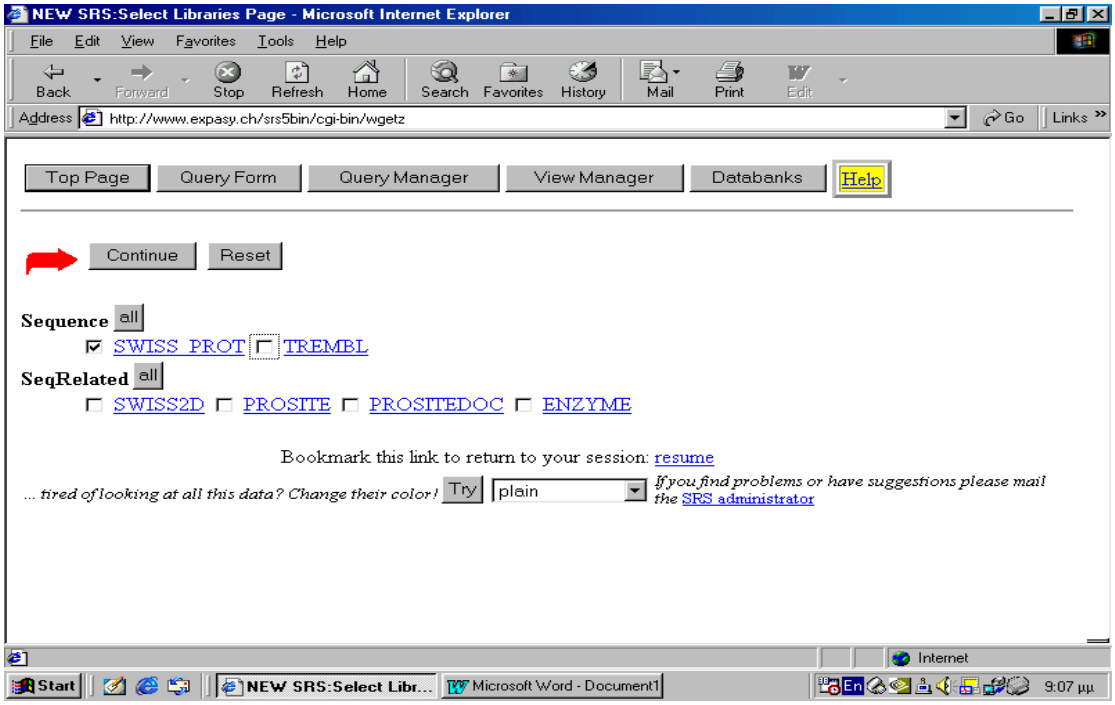


Μπορείς να μάθεις όλες τις λεπτομέρειες για τη βάση δεδομένων SWISS-PROT επιλέγοντας το “User Manual”, τότε θα εμφανισθεί η παρακάτω σελίδα.

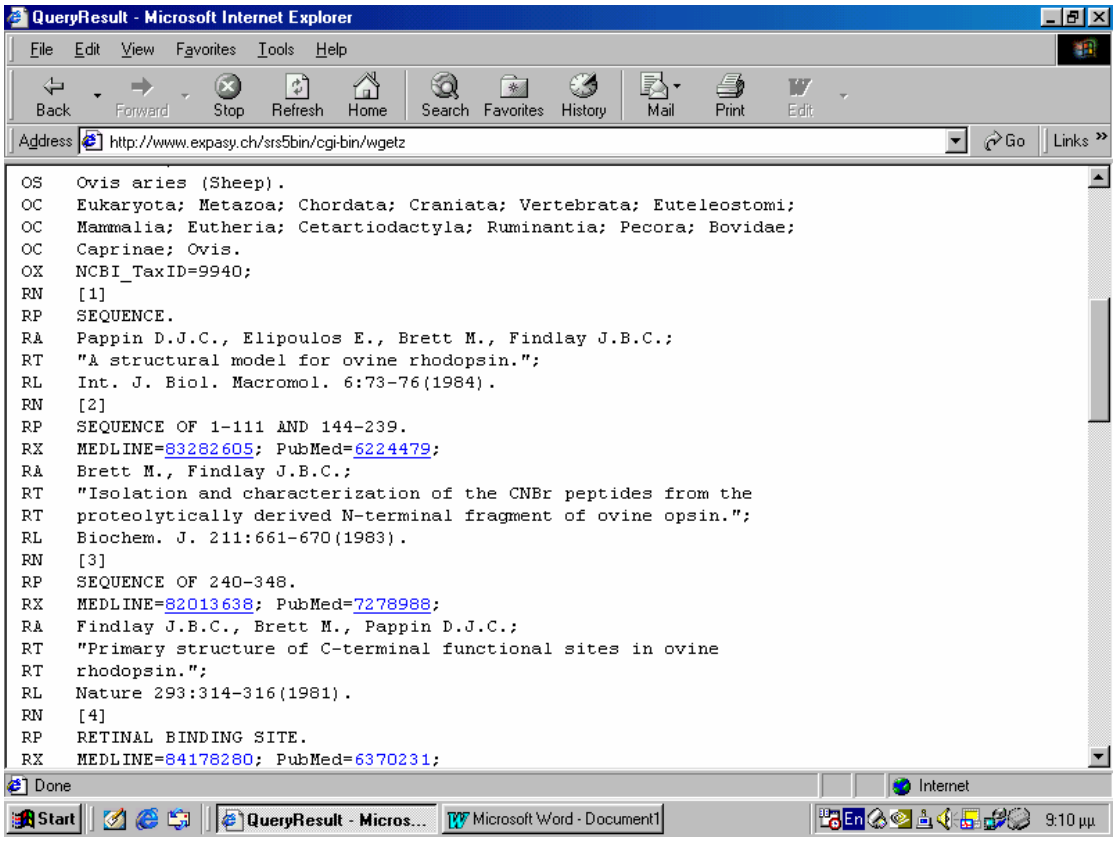


Για την ανάκτηση και την εξέταση της δομής μιας καταχώρησης στη SWISS-PROT, επιλέξτε SRS στο σημείο “Access to Swiss-Prot and TrEMBL” στην αρχική σελίδα της SWISS-PROT. Τότε θα εμφανιστεί η παρακάτω σελίδα. Το SRS είναι ένα ευκολόχρηστο σύστημα ανάκτησης, το οποίο επιτρέπει την εξαγωγή δεδομένων από τη βάση δεδομένων SWISS-PROT.

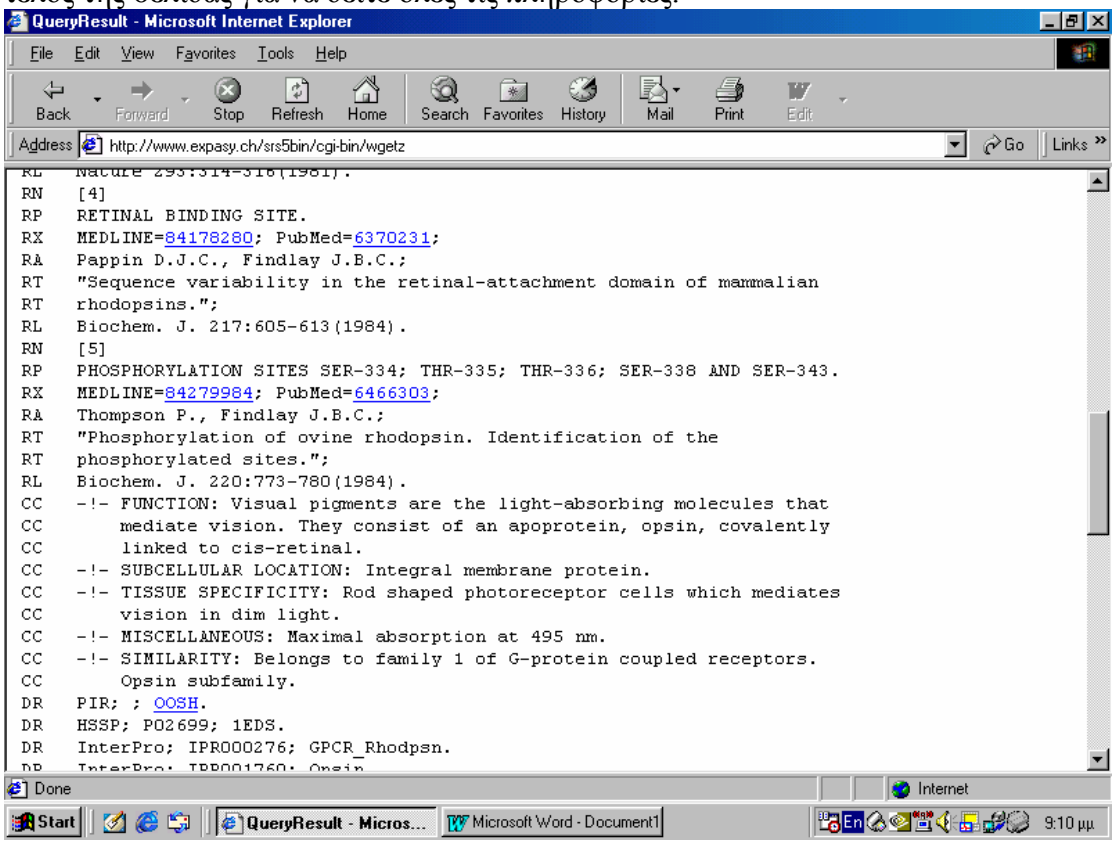
Μετά επιλέξτε “Start a new SRS session”. Στην επόμενη σελίδα επιλέξτε “ SWISS-PROT” και “Continue”..

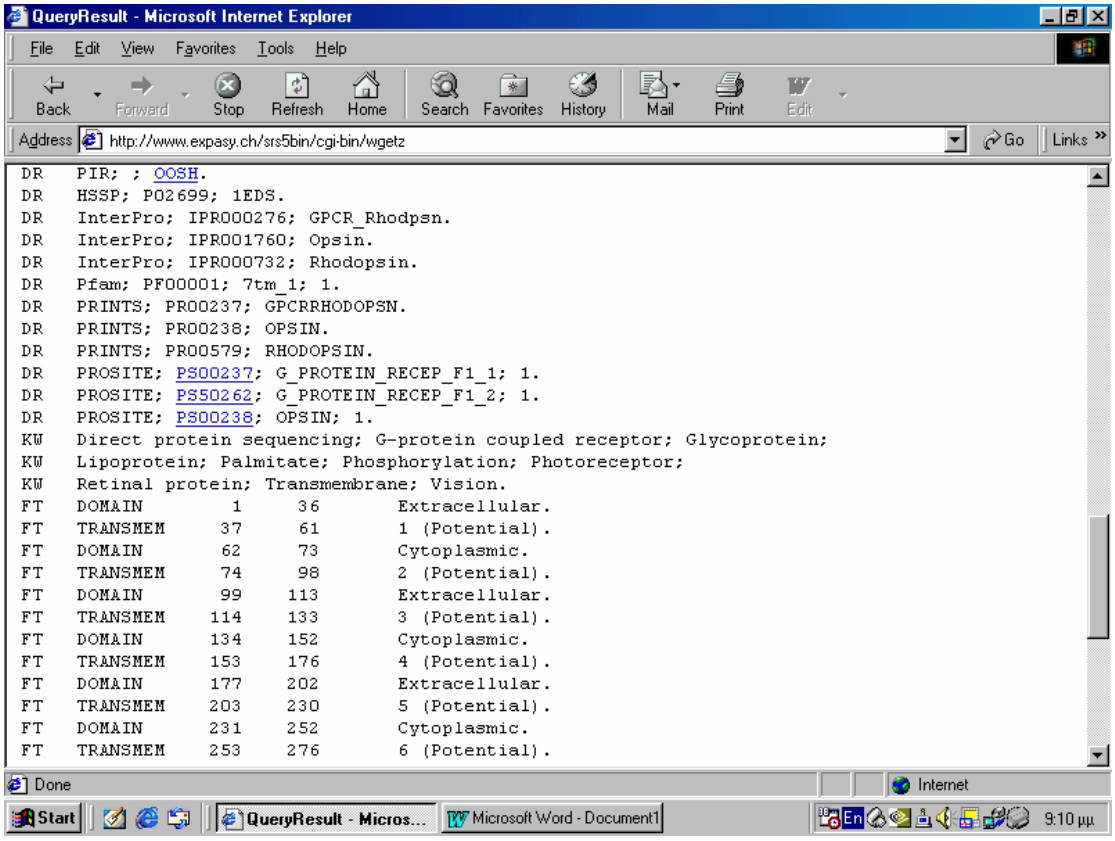


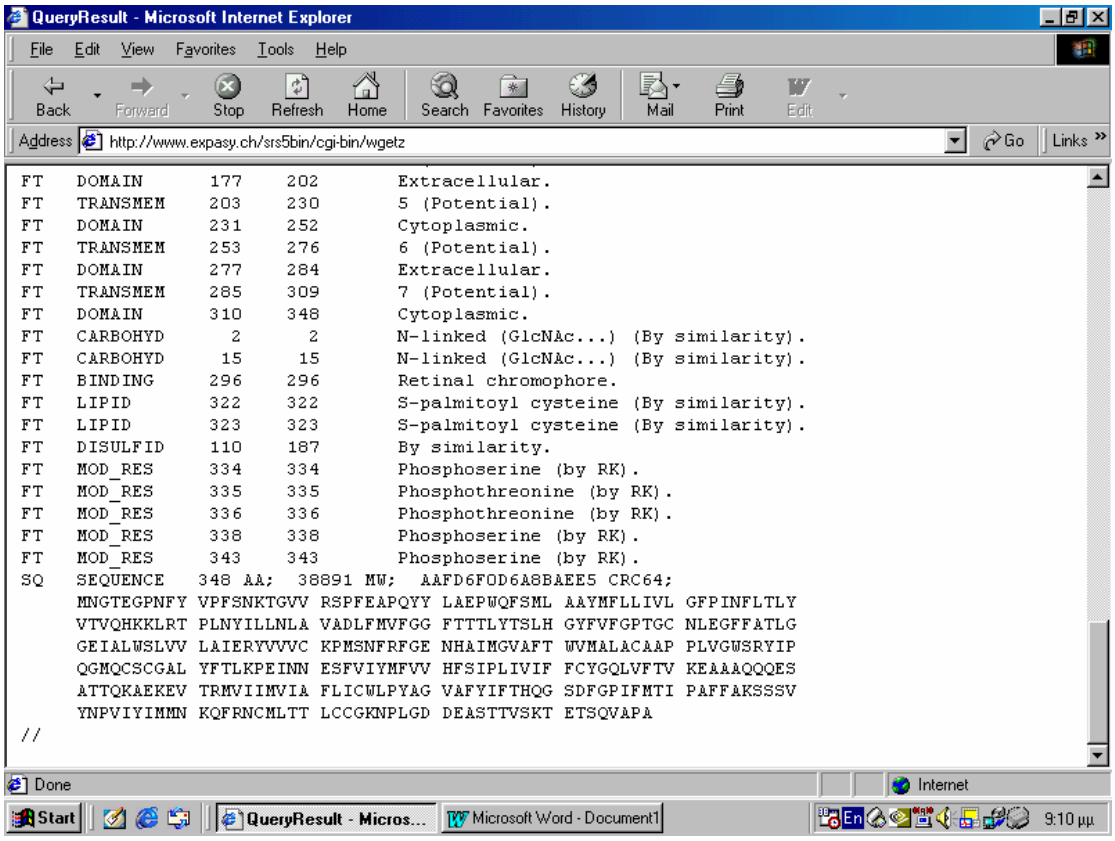
Για να δείτε τη δομή της καταχώρησης για την πρωτεΐνη OPSD_SHEEP, πληκτρολογήστε OPSD_SHEEP στο πεδίο δίπλα στο ID και επιλέξτε το “Do Query”.



Τότε θα εμφανιστεί ολόκληρη η εκχώρηση του OPSD_SHEEP. Προχωρήστε και στο τέλος της σελίδας για να δείτε όλες τις πληροφορίες.

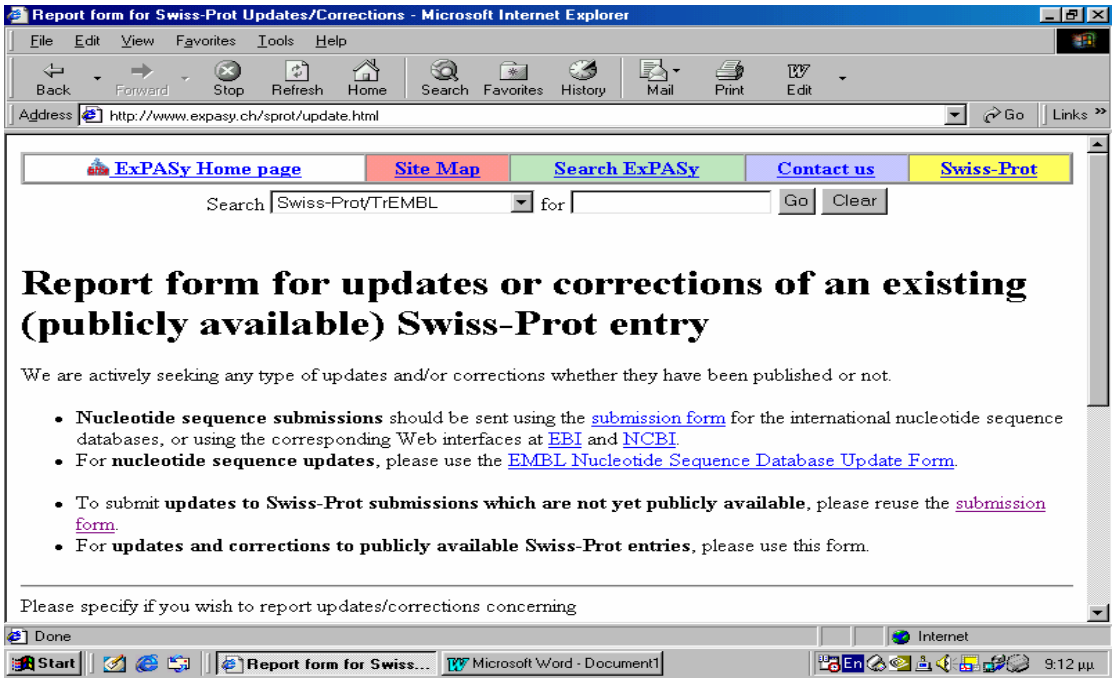



QueryResult - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://www.expasy.ch/srs5bin/cgi-bin/wgetz
Go Links
DR PIR; ; OOSH.
DR HSSP; P02699; 1EDS.
DR InterPro; IPR000276; GPCR_Rhodpsn.
DR InterPro; IPR001760; Opsin.
DR InterPro; IPR000732; Rhodopsin.
DR Pfam; PF00001; 7tm_1; 1.
DR PRINTS; PR00237; GPCRRHODOPSN.
DR PRINTS; PR00238; OPSIN.
DR PRINTS; PR00579; RHODOPSIN.
DR PROSITE; PS00237; G_PROTEIN_RECEP_F1_1; 1.
DR PROSITE; PS50262; G_PROTEIN_RECEP_F1_2; 1.
DR PROSITE; PS00238; OPSIN; 1.
KW Direct protein sequencing; G-protein coupled receptor; Glycoprotein;
KW Lipoprotein; Palmitate; Phosphorylation; Photoreceptor;
KW Retinal protein; Transmembrane; Vision.
FT DOMAIN 1 36 Extracellular.
FT TRANSMEM 37 61 1 (Potential).
FT DOMAIN 62 73 Cytoplasmic.
FT TRANSMEM 74 98 2 (Potential).
FT DOMAIN 99 113 Extracellular.
FT TRANSMEM 114 133 3 (Potential).
FT DOMAIN 134 152 Cytoplasmic.
FT TRANSMEM 153 176 4 (Potential).
FT DOMAIN 177 202 Extracellular.
FT TRANSMEM 203 230 5 (Potential).
FT DOMAIN 231 252 Cytoplasmic.
FT TRANSMEM 253 276 6 (Potential).
Done
Internet
Start
QueryResult - Micros...
Microsoft Word - Document1
9:10 μμ

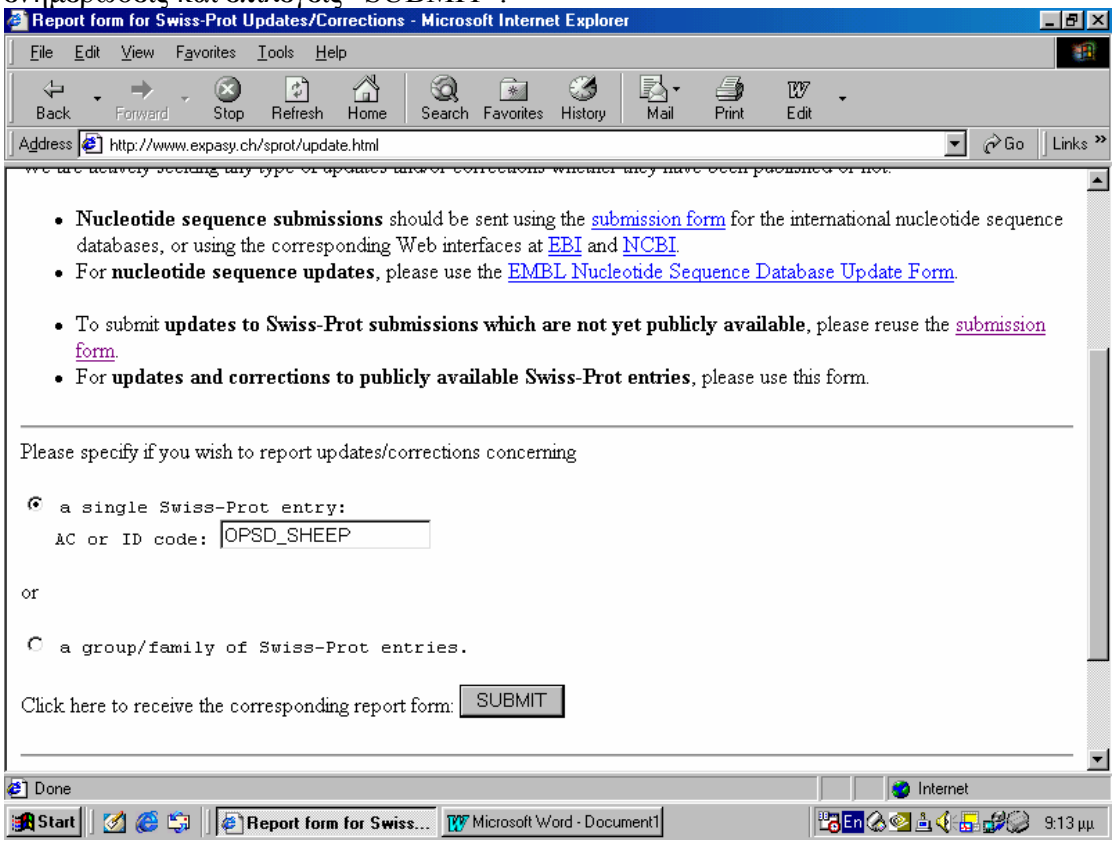

QueryResult - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://www.expasy.ch/srs5bin/cgi-bin/wgetz
Go Links
FT DOMAIN 177 202 Extracellular.
FT TRANSMEM 203 230 5 (Potential).
FT DOMAIN 231 252 Cytoplasmic.
FT TRANSMEM 253 276 6 (Potential).
FT DOMAIN 277 284 Extracellular.
FT TRANSMEM 285 309 7 (Potential).
FT DOMAIN 310 348 Cytoplasmic.
FT CARBOHYD 2 2 N-linked (GlcNAc...) (By similarity).
FT CARBOHYD 15 15 N-linked (GlcNAc...) (By similarity).
FT BINDING 296 296 Retinal chromophore.
FT LIPID 322 322 S-palmitoyl cysteine (By similarity).
FT LIPID 323 323 S-palmitoyl cysteine (By similarity).
FT DISULFID 110 187 By similarity.
FT MOD_RES 334 334 Phosphoserine (by RK).
FT MOD_RES 335 335 Phosphothreonine (by RK).
FT MOD_RES 336 336 Phosphothreonine (by RK).
FT MOD_RES 338 338 Phosphoserine (by RK).
FT MOD_RES 343 343 Phosphoserine (by RK).
SQ SEQUENCE 348 AA; 38891 MW; AAFD6F0D6A8BAEE5 CRC64;
MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIVL GFPINFLTLY
VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH GYFVFGPTGC NLEGFFATLG
GEIALWSLVV LAIERYVVVC KPMSNFRFGE NHAIMGVAFT WVMALACAAP PLVGWSRYIP
QGMQCSCGAL YFTLKPEINN ESFVIYMFVV HFSIPLIVIF FCYGQLVFTV KEAAAQQQES
ATTQKAEKEV TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKSSSV
YNPVIYIMMN KQFRNCMLTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
//
Done
Internet
Start
QueryResult - Micros...
Microsoft Word - Document1
9:10 μμ

Ενημέρωση δεδομένων στο SWISS-PROT

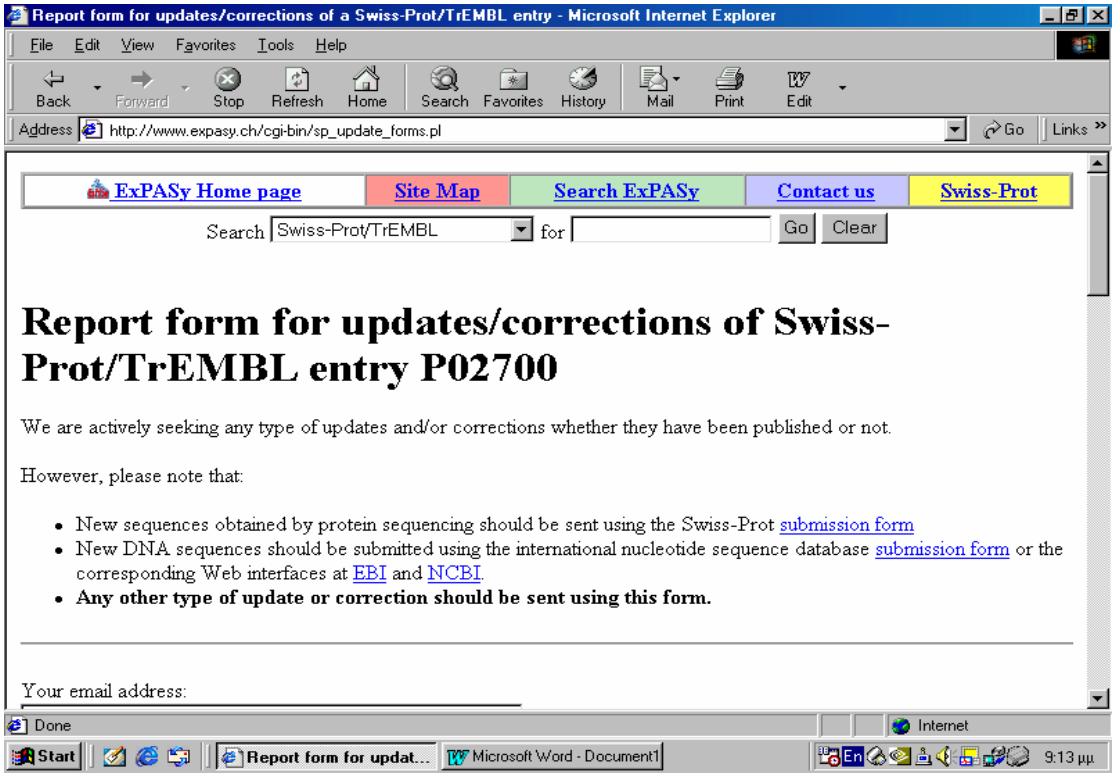
Αν θέλεις να ενημερώσεις τα δεδομένα που σχετίζονται με μια ήδη υπάρχον πρωτεΐνη, φορτώνεις την αρχική σελίδα του SWISS-PROT, προχωράς κάτω στη σελίδα μέχρι να φτάσεις στο σημείο "Documents and services" και επιλέγεις "Report form for updates or corrections".

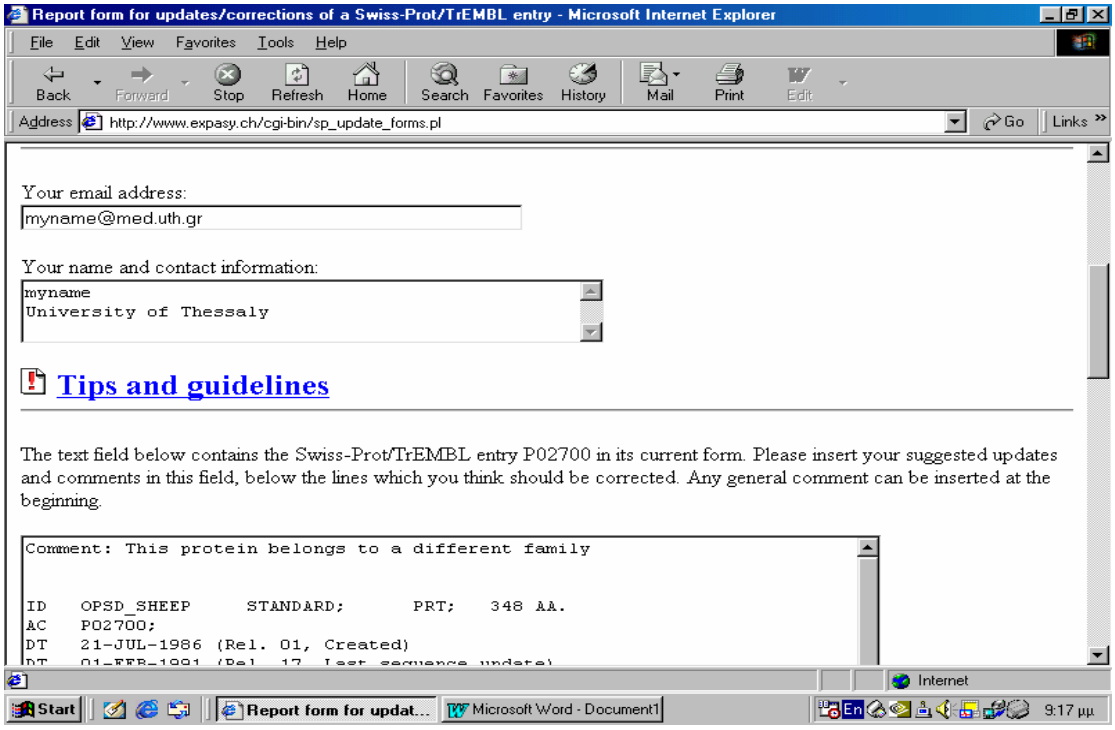


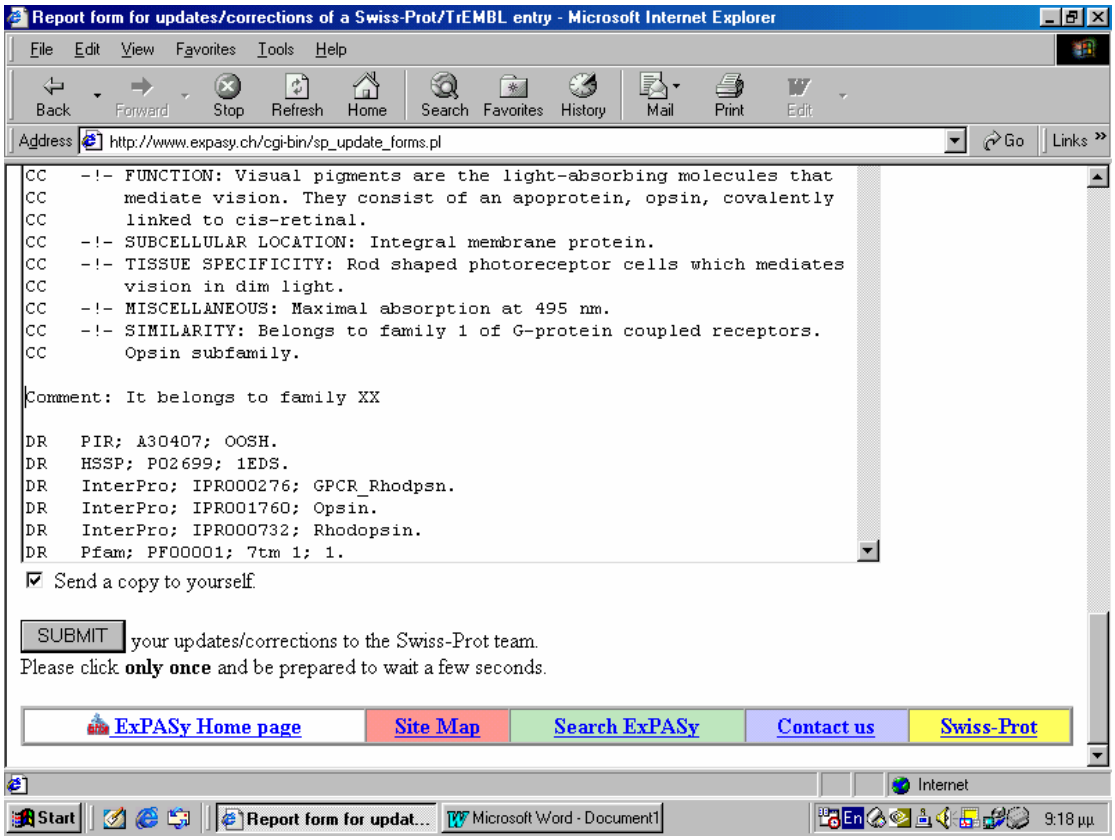
Μετά θα πρέπει να πληκτρολογήσεις το ID ή το AC της πρωτεΐνης που θέλεις να ενημερώσεις και επιλέγεις "SUBMIT".



Μετά η ιστοσελίδα με την καταχώρηση της πρωτεΐνης εμφανίζεται, πριν από οποιαδήποτε ενημέρωση θα πρέπει να δώσεις την ηλεκτρονική σου διεύθυνση (e-mail), το όνομά σου και τον φορέα σου. Στο παράθυρο της ιστοσελίδας, στην αρχή μπορείς να γράψεις ένα γενικό σχόλιο. Μετά μπορείς να αλλάξεις το περιεχόμενο της πρωτεΐνης με ευκρινή σχόλια.

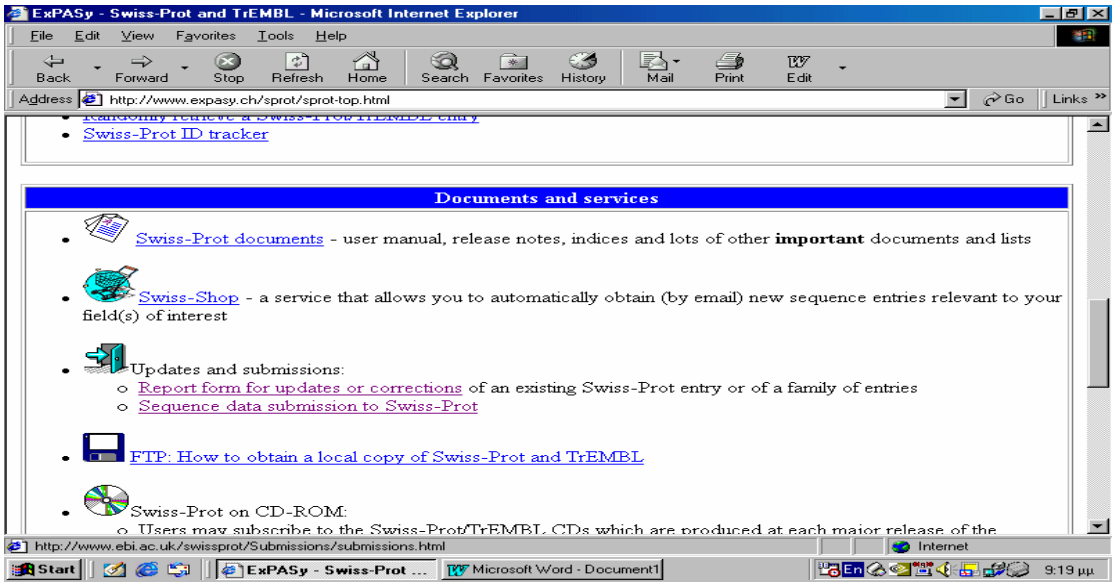




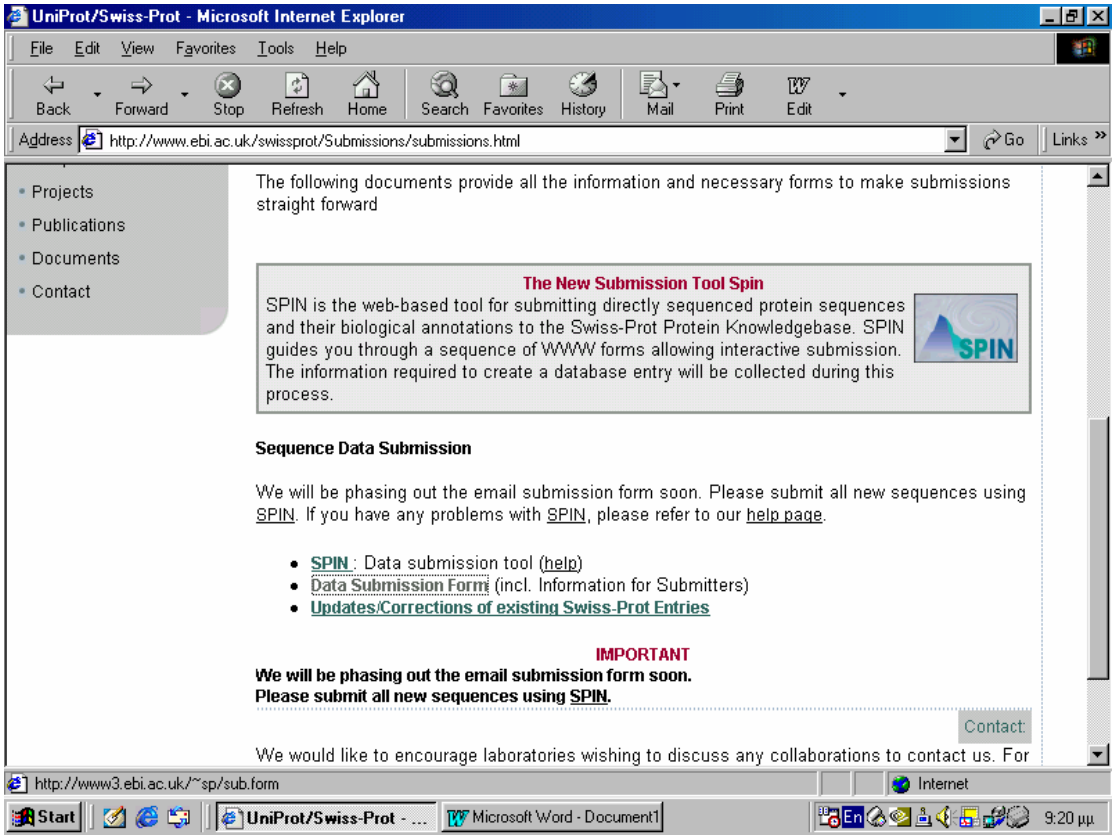

Report form for updates/corrections of a Swiss-Prot/TrEMBL entry - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://www.expasy.ch/cgi-bin/sp_update_forms.pl
Go Links
CC -!- FUNCTION: Visual pigments are the light-absorbing molecules that
CC mediate vision. They consist of an apoprotein, opsin, covalently
CC linked to cis-retinal.
CC -!- SUBCELLULAR LOCATION: Integral membrane protein.
CC -!- TISSUE SPECIFICITY: Rod shaped photoreceptor cells which mediates
CC vision in dim light.
CC -!- MISCELLANEOUS: Maximal absorption at 495 nm.
CC -!- SIMILARITY: Belongs to family 1 of G-protein coupled receptors.
CC Opsin subfamily.
Comment: It belongs to family XX
DR PIR; A30407; OOSH.
DR HSSP; P02699; 1EDS.
DR InterPro; IPR000276; GPCR_Rhodpsn.
DR InterPro; IPR001760; Opsin.
DR InterPro; IPR000732; Rhodopsin.
DR Pfam; PF00001; 7tm 1; 1.
Send a copy to yourself.
SUBMIT your updates/corrections to the Swiss-Prot team.
Please click only once and be prepared to wait a few seconds.
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Internet
Start
Report form for updat...
Microsoft Word - Document1
9:18 µµ

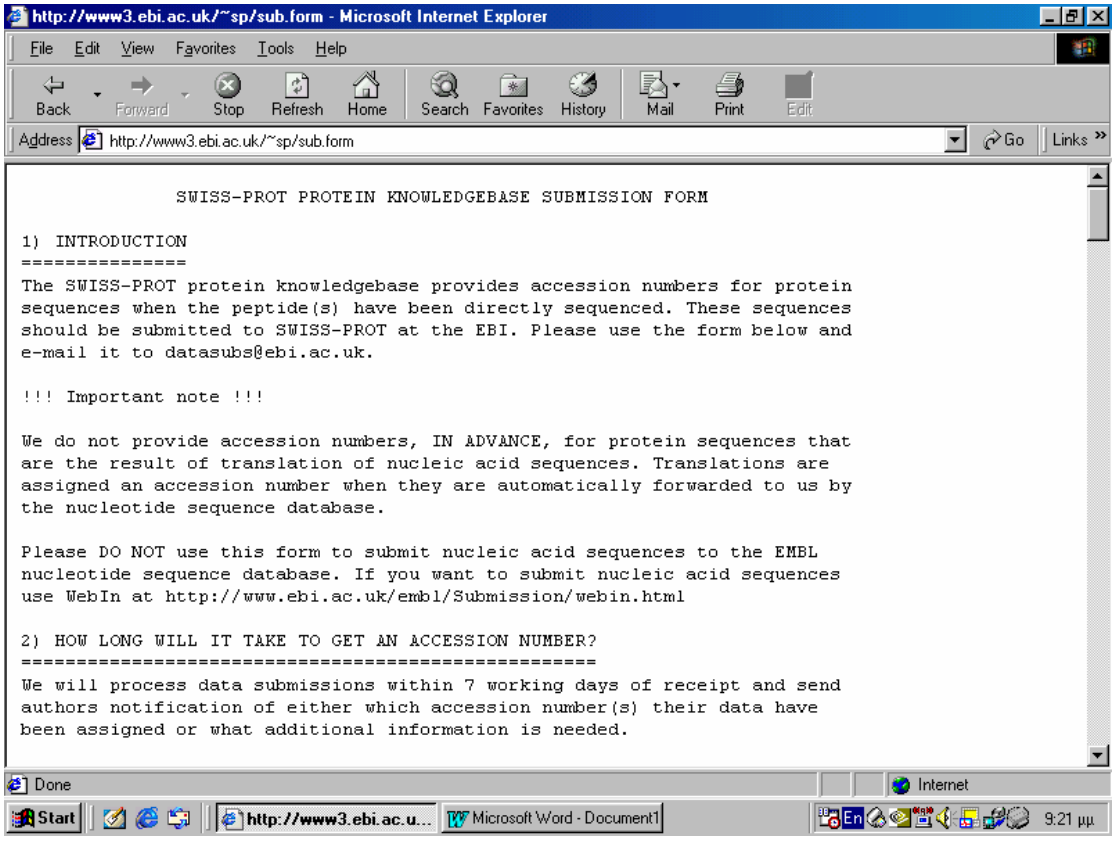
Καταγραφή μιας νέας πρωτεΐνης στο SWISS-PROT

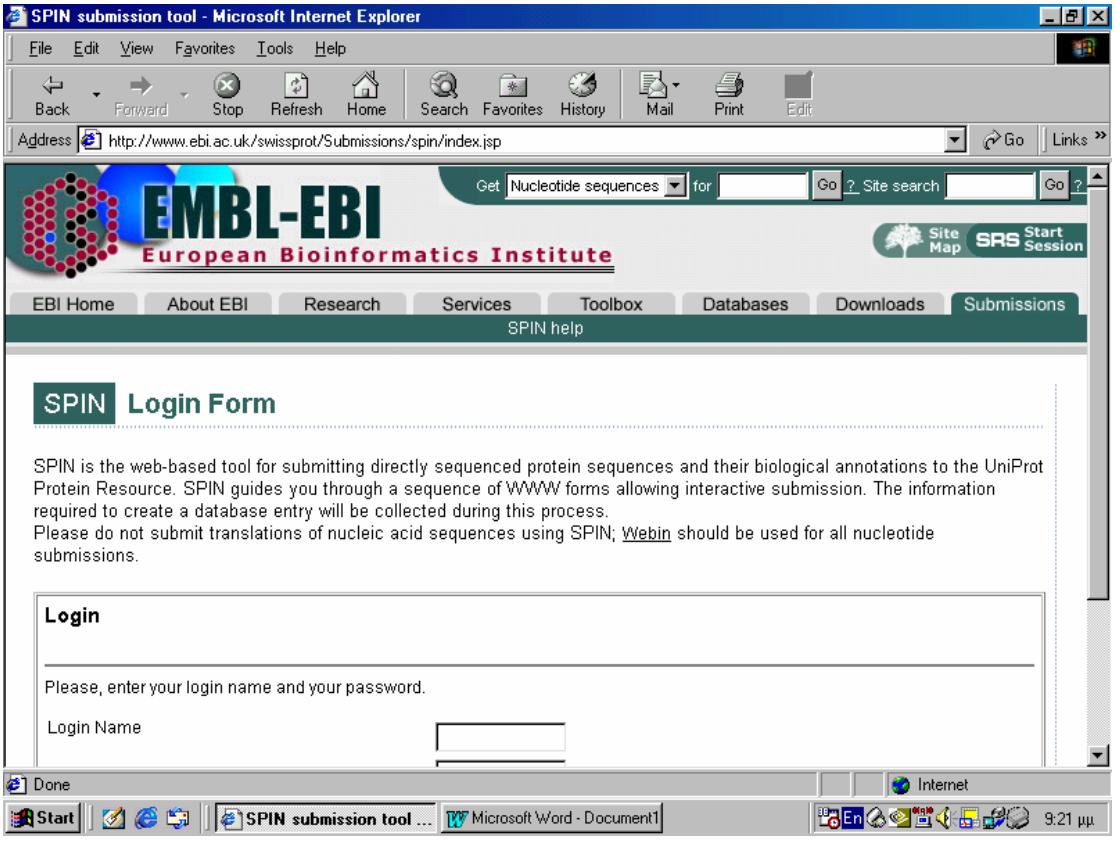
Αν θέλεις να καταγράψεις μια νέα πρωτινή με όλα τα απαραίτητα στοιχεία, φορτώνεις την αρχική σελίδα του SWISS-PROT, προχωράς κάτω μέχρι να φτάσεις στο σημείο "Documents and services" και κάνεις κλικ στην επιλογή "Sequence data submission to Swiss-prot".

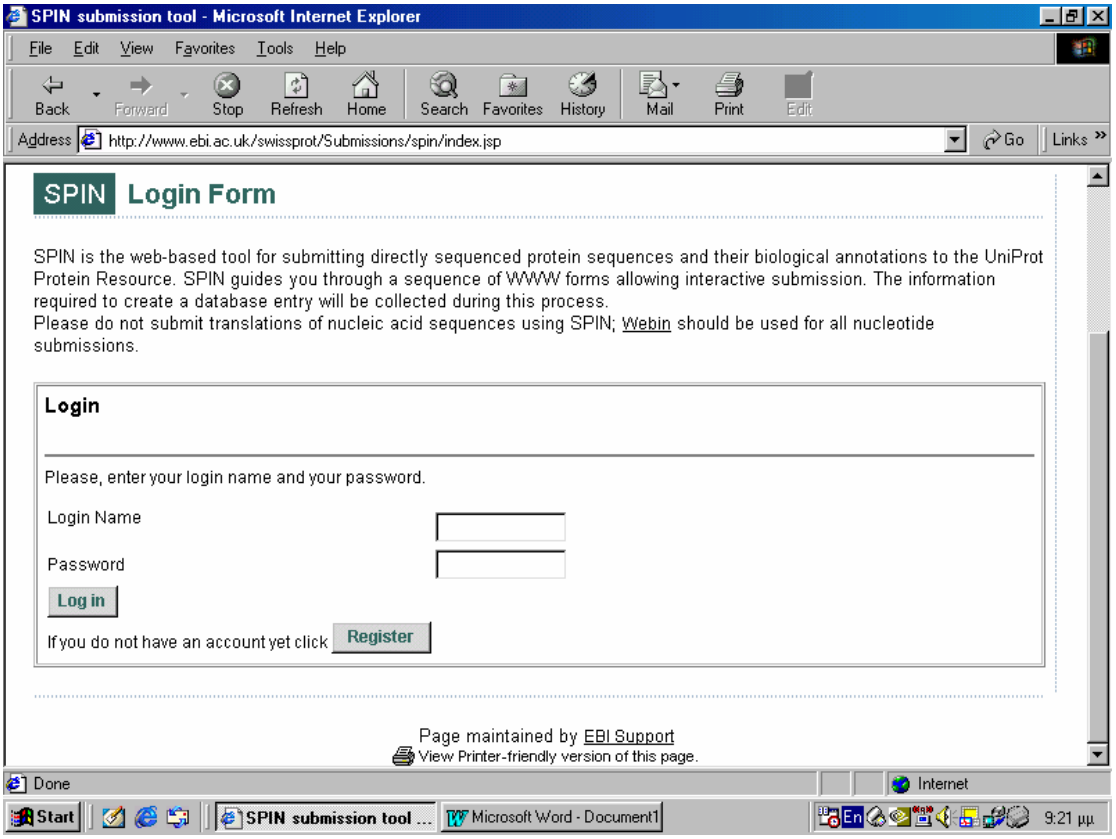


Τότε σε απευθύνει στην ιστοσελίδα του EMBL-EBI, μέσα στην ιστοσελίδα υποβολών του Swiss-Prot. Κάνοντας κλικ στο "SPIN", μπορείς να διαβάσεις τις οδηγίες σχετικά με την υποβολή μιας καταχώρησης. Επιλέγοντας το "Data Submission Form", σε βάζει σε μια σελίδα όπου σου ζητάει να γράψεις τα στοιχεία σου αν είσαι ήδη χρήστης αλλιώς να αποκτήσεις ένα λογαριασμό αν είσαι καινούριος χρήστης. Μετά την εγγραφή στοιχείων σου, μπορείς να υποβάλλεις την πρωτεΐνη σε μια δομή όμοια με την προηγούμενη (OPSD_SHEEP).


UniProt/Swiss-Prot - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://www.ebi.ac.uk/swissprot/Submissions/submissions.html
Go Links
Projects
Publications
Documents
Contact
The following documents provide all the information and necessary forms to make submissions straight forward
The New Submission Tool Spin
SPIN is the web-based tool for submitting directly sequenced protein sequences and their biological annotations to the Swiss-Prot Protein Knowledgebase. SPIN guides you through a sequence of WWW forms allowing interactive submission. The information required to create a database entry will be collected during this process.
SPIN
Sequence Data Submission
We will be phasing out the email submission form soon. Please submit all new sequences using SPIN. If you have any problems with SPIN, please refer to our help page.
SPIN : Data submission tool (help)
Data Submission Form (incl. Information for Submitters)
Updates/Corrections of existing Swiss-Prot Entries
IMPORTANT
We will be phasing out the email submission form soon.
Please submit all new sequences using SPIN.
Contact:
We would like to encourage laboratories wishing to discuss any collaborations to contact us. For
http://www3.ebi.ac.uk/~sp/sub.form
Internet
Start
UniProt/Swiss-Prot - ...
Microsoft Word - Document1
9:20 µµ


http://www3.ebi.ac.uk/~sp/sub.form - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://www3.ebi.ac.uk/~sp/sub.form
Go Links
SWISS-PROT PROTEIN KNOWLEDGEBASE SUBMISSION FORM
1) INTRODUCTION
The SWISS-PROT protein knowledgebase provides accession numbers for protein sequences when the peptide(s) have been directly sequenced. These sequences should be submitted to SWISS-PROT at the EBI. Please use the form below and e-mail it to datasubs@ebi.ac.uk.
!!! Important note !!!
We do not provide accession numbers, IN ADVANCE, for protein sequences that are the result of translation of nucleic acid sequences. Translations are assigned an accession number when they are automatically forwarded to us by the nucleotide sequence database.
Please DO NOT use this form to submit nucleic acid sequences to the EMBL nucleotide sequence database. If you want to submit nucleic acid sequences use WebIn at http://www.ebi.ac.uk/embl/Submission/webin.html
2) HOW LONG WILL IT TAKE TO GET AN ACCESSION NUMBER?
We will process data submissions within 7 working days of receipt and send authors notification of either which accession number(s) their data have been assigned or what additional information is needed.
Done
Internet
Start
http://www3.ebi.ac.u...
Microsoft Word - Document1
9:21 µµ


SPIN submission tool - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://www.ebi.ac.uk/swissprot/Submissions/spin/index.jsp
Go Links
EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go ? Site search Go ?
Site Map SRS Start Session
EBI Home About EBI Research Services Toolbox Databases Downloads Submissions
SPIN help
SPIN Login Form
SPIN is the web-based tool for submitting directly sequenced protein sequences and their biological annotations to the UniProt Protein Resource. SPIN guides you through a sequence of WWW forms allowing interactive submission. The information required to create a database entry will be collected during this process.
Please do not submit translations of nucleic acid sequences using SPIN; Webin should be used for all nucleotide submissions.
Login
Please, enter your login name and your password.
Login Name
Done
Internet
Start
SPIN submission tool ...
Microsoft Word - Document1
9:21 μμ


SPIN submission tool - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://www.ebi.ac.uk/swissprot/Submissions/spin/index.jsp
Go Links
SPIN Login Form
SPIN is the web-based tool for submitting directly sequenced protein sequences and their biological annotations to the UniProt Protein Resource. SPIN guides you through a sequence of WWW forms allowing interactive submission. The information required to create a database entry will be collected during this process.
Please do not submit translations of nucleic acid sequences using SPIN; Webin should be used for all nucleotide submissions.
Login
Please, enter your login name and your password.
Login Name
Password
Log in
If you do not have an account yet click Register
Page maintained by EBI Support
View Printer-friendly version of this page.
Done
Internet
Start
SPIN submission tool ...
Microsoft Word - Document1
9:21 μμ

Δευτεροταγείς βάσεις δεδομένων (Secondary databases)

Οι δευτεροταγείς (pattern) βάσεις δεδομένων (ΒΔ) περιέχουν τα αποτελέσματα από τις αναλύσεις των αλληλουχιών που βρίσκονται στις κύριες πηγές πληροφόρησης. Επειδή υπάρχουν πολλές διαφορετικές κύριες βάσεις δεδομένων και διαφορετικοί τρόποι ανάλυσης αλληλουχιών από πρωτεΐνες, οι πληροφορίες που είναι αποθηκευμένες σε κάθε δευτεροταγή βάση δεδομένων και καθώς και η οργάνωσή τους είναι διαφορετικές.

Η SWISS-PROT αποτελεί τη βάση για πολλές δευτεροταγείς βάσεις δεδομένων. Οι πιο σημαντικές δευτεροταγείς βάσεις δεδομένων είναι οι ακόλουθες:

Δευτεροταγής ΒΔ	**Κύρια πηγή**	**Αποθηκευμένη Πληροφορία**
PROSITE	SWISS-PROT	Regular expressions (patterns)
Profiles (profiles)	SWISS-PROT	Weighted matrices
PRINTS (fingerprints)	OWL	Aligned motifs
Pfam	SWISS-PROT	Hidden Markov Models (HMMs)
BLOCKS	PROSITE/PRINTS	Aligned motifs (blocks)
IDENTIFY	BLOCKS/PRINTS	Fuzzy regular expressions (patterns)

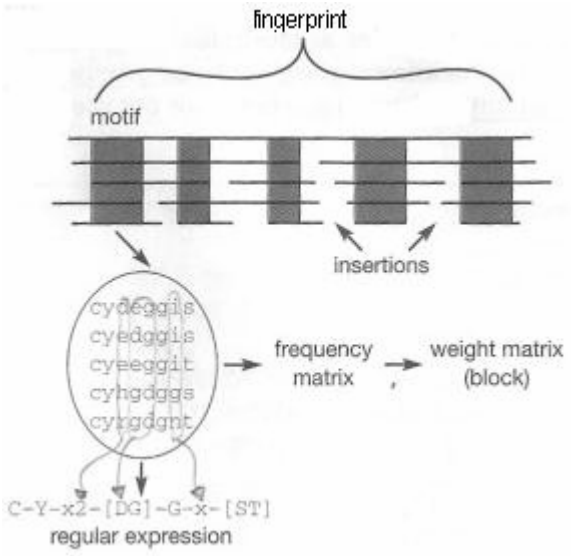
Ο τύπος πληροφορίας που είναι αποθηκευμένος σε κάθε δευτεροταγή βάση δεδομένων διαφέρει. Οι δευτεροταγείς βάσεις δεδομένων αποθηκεύουν μοτίβα (συγκεκριμένοι συνδυασμοί α-ελίκων και β-φύλλων, οι οποίοι παρουσιάζονται επανειλημμένως σε πολλές, ασύνδετες μεταξύ τους, πρωτεΐνες), τα οποία είναι συντηρημένες περιοχές με μικρή διακύμανση μεταξύ των αλληλουχιών και προκύπτουν ως αποτέλεσμα πολλαπλών αντιστοιχιών.

Οι συντηρημένες αυτές περιοχές έχουν μήκος 10-20 αμινοξέα και συνήθως αντιστοιχούν στα πιο σημαντικά στοιχεία που καθορίζουν τη δομή και τη λειτουργικότητα της πρωτεΐνης. Τα μοτίβα χρησιμοποιούνται για τη διάγνωση μελών μιας οικογένειας μέσα από την εφαρμογή μιας σειράς τεχνικών για την ανάλυση αλληλουχιών.

Τα μοτίβα χρησιμοποιούνται για τη δημιουργία διαγνωστικών μοντέλων (patterns) για συγκεκριμένες οικογένειες πρωτεϊνών: μία άγνωστη αλληλουχία αναζητείται και συγκρίνεται με μία βιβλιοθήκη από τέτοια μοντέλα (μοτίβα) προκειμένου να προσδιοριστεί εάν περιλαμβάνει ή όχι κάποιο μοτίβο. Στη συνέχεια, και με βάση το μοτίβο αυτό, η αλληλουχία αποδίδεται σε μία γνωστή οικογένεια.

Τα μοτίβα – motifs (ή αλλιώς τεμάχια – blocks, ή τμήματα – segments, ή χαρακτηριστικά – features), αντανακλούν το βιολογικό ρόλο της πρωτεΐνης, δηλαδή προσδιορίζουν τη δομή ή τη λειτουργία της πρωτεΐνης.

Σε περίπτωση που η δομή και η λειτουργία της οικογένειας είναι γνωστές, η αναζήτηση των δευτεροταγών (pattern) βάσεων δεδομένων παρέχει πληροφορίες σχετικά με τη βιολογική λειτουργία της πρωτεΐνης. Καθώς οι δευτεροταγείς βάσεις δεδομένων συγκροτούνται από πολλαπλές πληροφορίες αλληλουχιών, οι αναζητήσεις σε αυτές παρέχουν καλύτερα αποτελέσματα σχετικά με μακρινές σχέσεις πρωτεϊνών, συγκριτικά με τα αποτελέσματα αντίστοιχων αναζητήσεων σε κύριες βάσεις δεδομένων.



Regular expression (pattern) είναι μία σύντομη περιγραφή ενός μοτίβου: το x συμβολίζει ένα οποιοδήποτε αμινοξύ και στις παρενθέσεις είναι τα αμινοξέα που επιτρέπονται σε αυτή τη θέση.

Fingerprint (ή αλλιώς υπογραφή - signature) είναι ένα σύνολο από μοτίβα. Όλες οι πληροφορίες για τα αμινοξέα διατηρούνται με την μορφή πινάκων συχνοτήτων.

Block είναι ένα fingerprint όπου στον πίνακα συχνοτήτων προστίθεται μία βαθμολόγηση.

Profile είναι ένα block που χρησιμοποιεί τις πληροφορίες από όλη την αντιστοίχηση, δηλαδή συμπεριλαμβάνει και τα κενά.

Hidden Markov Models είναι μοντέλα πιθανοτήτων που παράγονται από τα profiles.

PROSITE

Η PROSITE διατηρείται στο Swiss Institute of Bioinformatics. Στην PROSITE, οι οικογένειες των πρωτεϊνών μπορούν να χαρακτηριστούν από το μοναδικό, πιο διατηρημένο μοτίβο το οποίο εντοπίζεται σε μία πολλαπλή αντιστοίχηση γνωστών ομολόγων. Τα μοτίβα αυτά περιέχουν κωδικοποιημένές βιολογικές λειτουργίες – κλειδιά: ενεργές περιοχές ενζύμων, θέσεις στις οποίες ο συνδέτης συνδέεται με το υπόστρωμα, κ.λ.π.
Με αναζήτηση στην PROSITE μπορεί να προσδιοριστεί η οικογένεια πρωτεϊνών στην οποία ανήκει μία νέα αλληλουχία. Υπάρχει επίσης η δυνατότητα να προσδιοριστούν οι περιοχές τις οποίες περιέχει η αλληλουχία αυτή.

Στην PROSITE, τα μοτίβα κωδικοποιούνται σαν κανονικές εκφράσεις (ή μοντέλα - patterns).

Η PROSITE λειτουργεί ως εξής: Τα μοντέλα (patterns) προέρχονται από την κατασκευή μίας πολλαπλής αντιστοίχησης (multiple alignment) και από τον έλεγχο με το χέρι και αναγνώριση των διατηρημένων περιοχών (conserved regions). Στη συνέχεια, χρησιμοποιώντας ένα ήδη αναγνωρισμένο μοντέλο, πραγματοποιείται μία

αναζήτηση στην SWISS-PROT προκειμένου να αναγνωριστούν οι πρωτεΐνες με παρόμοια μοντέλα, και έπειτα ελέγχεται η επίδοση / αξία του μοντέλου: πρέπει να υπάρχουν μόνο σωστές αντιστοιχήσεις / ταιριάσματα (αληθινές – θετικές) και καθόλου λανθασμένες αντιστοιχήσεις / ταιριάσματα (λανθασμένες – θετικές). Σε περίπτωση που το μοντέλο παράγει πολλές λανθασμένες – θετικές, τότε προτιμάται να χρησιμοποιηθεί κάποιο άλλο μοντέλο.

Δομή καταχωρήσεων της PROSITE

Υπάρχουν δύο τύποι καταχώρησης στη PROSITE: ο ένας είναι ένα αρχείο δεδομένων που κατασκευάζεται με παρόμοιο τρόπο όπως στη SWISS-PROT και ο άλλος είναι ένα αρχείο κειμένου.

Το αρχείο δεδομένων έχει την ακόλουθη μορφή:

```
ID   OPSIN; PATTERN.
AC   PS00238;
DT   APR-1990 (CREATED); DEC-2001 (DATA UPDATE); AUG-2004 (INFO UPDATE).
DE   Visual pigments (opsins) retinal binding site.
PA   [LIVMFWAC]-[PSGAC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-[STACP]-x(2)-[DENF]-
PA   [AP]-x(2)-[IY].
NR   /RELEASE=44.2,157002;
NR   /TOTAL=197(196); /POSITIVE=192(191); /UNKNOWN=0(0); /FALSE_POS=5(5);
NR   /FALSE_NEG=1; /PARTIAL=4;
CC   /TAXO-RANGE=??E??; /MAX-REPEAT=2;
CC   /SITE=5,retinal;
DR   Q9H1Y3, OPN3_HUMAN, T; Q9WUK7, OPN3_MOUSE, T; Q9UHM6, OPN4_HUMAN, T;
DR   Q9QXZ9, OPN4_MOUSE, T; P22269, OPS1_CALVI, T; P06002, OPS1_DROME, T;
DR   P28678, OPS1_DROPS, T; Q25157, OPS1_HEMSA, T; P35360, OPS1_LIMPO, T;
DR   O15973, OPS1_PATYE, T; Q94741, OPS1_SCHGR, T; P08099, OPS2_DROME, T;
DR   P28679, OPS2_DROPS, T; Q25158, OPS2_HEMSA, T; P35361, OPS2_LIMPO, T;
DR   O15974, OPS2_PATYE, T; Q26495, OPS2_SCHGR, T; P04950, OPS3_DROME, T;
DR   P28680, OPS3_DROPS, T; P08255, OPS4_DROME, T; P29404, OPS4_DROPS, T;
DR   P17646, OPS4_DROVI, T; P91657, OPS5_DROME, T; O01668, OPS6_DROME, T;
DR   P51471, OPSB_ANOCA, T; P90680, OPSB_APIME, T; P51472, OPSB_ASTFA, T;
DR   P51490, OPSB_BOVIN, T; Q9W6A8, OPSB_BRARE, T; P32310, OPSB_CARAU, T;
DR   P28682, OPSB_CHICK, T; O13227, OPSB_CONCO, T; P35357, OPSB_GECGE, T;
DR   P03999, OPSB_HUMAN, T; P51491, OPSB_MOUSE, T; P87365, OPSB_ORYLA, T;
DR   P60573, OPSB_PANPA, T; P60015, OPSB_PANTR, T; Q63652, OPSB_RAT  , T;
DR   O13092, OPSB_SAIBB, T; O42294, OPSD_ABYKO, T; P52202, OPSD_ALLMI, T;
DR   Q90245, OPSD_AMBTI, T; Q90214, OPSD_ANGAN, T; P41591, OPSD_ANOCA, T;
DR   Q17053, OPSD_APIME, T; P41590, OPSD_ASTFA, T; Q9YGZ1, OPSD_ATHBO, T;
DR   O42300, OPSD_BATMU, T; O42301, OPSD_BATNI, T; P02699, OPSD_BOVIN, T;
DR   P35359, OPSD_BRARE, T; P56514, OPSD_BUFBU, T; P56515, OPSD_BUFMA, T;
DR   Q17292, OPSD_CAMAB, T; O18312, OPSD_CAMHU, T; O16017, OPSD_CAMLU, T;
DR   O18315, OPSD_CAMMA, T; O16018, OPSD_CAMSC, T; P32308, OPSD_CANFA, T;
DR   P32309, OPSD_CARAU, T; Q17296, OPSD_CATBO, T; Q9YGZ8, OPSD_CHELB, T;
DR   P22328, OPSD_CHICK, T; O42327, OPSD_COMDY, T; Q90305, OPSD_CORAU, T;
DR   O42307, OPSD_COTBO, T; O42328, OPSD_COTGR, T; O42330, OPSD_COTIN, T;
DR   Q90373, OPSD_COTKE, T; P28681, OPSD_CRIGR, T; P51488, OPSD_CYPCA, T;
DR   O62791, OPSD_DELDE, T; Q9YGZ4, OPSD_DICLA, T; Q9YH05, OPSD_DIPAN, T;
DR   Q9YH04, OPSD_DIPVU, T; O93441, OPSD_GALML, T; P79756, OPSD_GAMAF, T;
DR   O62792, OPSD_GLOME, T; Q9YGZ2, OPSD_GOBNI, T; P08100, OPSD_HUMAN, T;
DR   O42268, OPSD_ICTPU, T; P22671, OPSD_LAMJA, T; O42427, OPSD_LIMBE, T;
DR   O42431, OPSD_LIMPA, T; Q9YH00, OPSD_LITMO, T; Q9YGZ6, OPSD_LIZAU, T;
DR   Q9YGZ7, OPSD_LIZSA, T; P24603, OPSD_LOLFO, T; Q17094, OPSD_LOLSU, T;
DR   Q28886, OPSD_MACFA, T; O62793, OPSD_MESBI, T; P15409, OPSD_MOUSE, T;
DR   Q9YGZ9, OPSD_MUGCE, T; Q9YH01, OPSD_MULSU, T; P79798, OPSD_MYRBE, T;
DR   P79807, OPSD_MYRVI, T; P79808, OPSD_NEOAR, T; P79809, OPSD_NEOAU, T;
DR   P79812, OPSD_NEOSA, T; P09241, OPSD_OCTDO, T; O18481, OPSD_ORCAU, T;
DR   O16019, OPSD_ORCVI, T; P87369, OPSD_ORYLA, T; O42452, OPSD_PARKN, T;
DR   Q98980, OPSD_PETMA, T; O62795, OPSD_PHOGR, T; O62794, OPSD_PHOVI, T;
DR   O18766, OPSD_PIG  , T; P79848, OPSD_POERE, T; P35403, OPSD_POMMI, T;
DR   P35356, OPSD_PROCL, T; O42451, OPSD_PROJE, T; O16020, OPSD_PROML, T;
DR   O18485, OPSD_PROOR, T; O18486, OPSD_PROSE, T; P49912, OPSD_RABIT, T;
DR   P79863, OPSD_RAJER, T; P51470, OPSD_RANCA, T; P31355, OPSD_RANPI, T;
DR   P56516, OPSD_RANTE, T; P51489, OPSD_RAT  , T; Q9YGZ3, OPSD_SALPV, T;
DR   P79898, OPSD_SARDI, T; P79901, OPSD_SARMI, T; Q9YGZ0, OPSD_SARPI, T;
DR   P79902, OPSD_SARPU, T; Q9YH03, OPSD_SARSL, T; P79903, OPSD_SARSP, T;
DR   P79911, OPSD_SARTI, T; P79914, OPSD_SARXA, T; O93459, OPSD_SCYCA, T;
DR   O16005, OPSD_SEPOF, T; P02700, OPSD_SHEEP, T; Q8HY69, OPSD_SMICR, T;
DR   Q9YGZ5, OPSD_SOLSO, T; Q9YH02, OPSD_SPAAU, T; P35362, OPSD_SPHSP, T;
DR   O42466, OPSD_TAUBU, T; Q9DGG4, OPSD_TETNG, T; P31356, OPSD_TODPA, T;
DR   O62796, OPSD_TRIMA, T; O62798, OPSD_TURTR, T; P29403, OPSD_XENLA, T;
DR   O42604, OPSD_ZEUFA, T; Q9YGY9, OPSD_ZOSOP, T; Q90215, OPSF_ANGAN, T;
DR   P22330, OPSG_ASTFA, T; Q9W6A5, OPSG_BRARE, T; P32311, OPSG_CARAU, T;
DR   Q9R024, OPSG_CAVPO, T; P28683, OPSG_CHICK, T; P35358, OPSG_GECGE, T;
DR   P04001, OPSG_HUMAN, T; O35599, OPSG_MOUSE, T; P87366, OPSG_ORYLA, T;
DR   O18910, OPSG_RABIT, T; O35476, OPSG_RAT  , T; O35478, OPSG_SCICA, T;
DR   P22331, OPSH_ASTFA, T; Q9W6A6, OPSH_BRARE, T; P32312, OPSH_CARAU, T;
DR   P51474, OPSI_ASTFA, T; P34989, OPSL_CALJA, T; O13018, OPSO_SALSA, T;
DR   P51475, OPSP_CHICK, T; P51476, OPSP_COLLI, T; O42266, OPSP_ICTPU, T;
DR   O42490, OPSP_PETMA, T; P41592, OPSR_ANOCA, T; P22332, OPSR_ASTFA, T;
DR   Q9W6A7, OPSR_BRARE, T; Q95170, OPSR_CAPHI, T; P32313, OPSR_CARAU, T;
DR   P22329, OPSR_CHICK, T; O18913, OPSR_FELCA, T; P04000, OPSR_HUMAN, T;
DR   P87367, OPSR_ORYLA, T; O12948, OPSR_XENLA, T; Q9W6A9, OPSU_BRARE, T;
DR   Q90309, OPSU_CARAU, T; O57605, OPSU_MELUD, T; O61303, OPSV_APIME, T;
DR   P28684, OPSV_CHICK, T; P87368, OPSV_ORYLA, T; P51473, OPSV_XENLA, T;
DR   O14718, OPSX_HUMAN, T; O35214, OPSX_MOUSE, T; P23820, REIS_TODPA, T;
DR   P47803, RGR_BOVIN , T; P47804, RGR_HUMAN , T;
DR   P17645, OPS3_DROVI, P; O18911, OPSG_ODOVI, P; O18914, OPSR_CANFA, P;
DR   O18912, OPSR_HORSE, P;
```

```
DR   Q9Z2B3, RGR_MOUSE , N;
DR   Q6MLD2, GUAA_BDEBA, F; Q9CL24, OADB_PASMU, F; P22056, POLS_ONNVG, F;
DR   Q99NF8, RP17_MOUSE, F; P09009, TERM_BPPRD, F;
3D   1BOJ; 1BOK; 1F88; 1GZM; 1HZX; 1JFP; 1KPN; 1KPW; 1KPX; 1L9H; 1LN6;
DO   PDOC00211;
//
```

ID είναι ένα προσδιοριστικό, ένα ακρώνυμο για την οικογένεια και δείχνει το είδος του διαχωρισμού.
PATTERN σημαίνει ότι έχει χρησιμοποιηθεί μία regular expression.
AC είναι ένας αριθμός πρόσβασης (accession number).
Η γραμμή DE περιέχει περιγραφή της οικογένειας.
Η γραμμή DE περιέχει περιγραφή της οικογένειας.
Η γραμμή NR δίνει τεχνικές λεπτομέρειες και την διαγνωστική επίδοση / αξία του pattern. Έχουν βρεθεί 197 όμοιες αλληλουχίες με 5 λανθασμένες ομοιότητες.
Οι γραμμές CC παρέχουν πληροφορίες σχετικά με την ταξινόμηση της οικογένειας, το μέγιστο αριθμό επαναλήψεων του pattern σε μία αναζήτηση, κ.λ.π.
Μετά τα σχόλια ακολουθούν οι γραμμές DR με λίστες με αριθμούς πρόσβασης και προσδιοριστικούς κωδικούς της SWISS-PROT, όπου υπάρχει μία ένδειξη αν το pattern είναι πραγματικό (T), πιθανό (P), λανθασμένα θετικό (F), ή λανθασμένα αρνητικό (N).

Το αρχείο κειμένου έχει ελεύθερη μορφή:

```
{PDOC00211}
{PS00238; OPSIN}
{BEGIN}
**********************************************
* Visual pigments (opsins) retinal binding site *
**********************************************

Visual pigments [1,2] are the light-absorbing  molecules that  mediate vision.
They consist of  an apoprotein, opsin,  covalently  linked  to the chromophore
cis-retinal.  Vision is  effected through  the absorption of a  photon by cis-
retinal  which is isomerized to  trans-retinal.  This isomerization leads to a
change  of conformation  of the protein. Opsins are integral membrane proteins
with  seven transmembrane regions that belong to family 1 of G-protein coupled
receptors (see <PDOC00210>).

In vertebrates four different pigments are generally found.   Rod cells, which
mediate vision in dim light, contain the pigment rhodopsin.  Cone cells, which
function in bright light, are responsible  for  color vision and contain three
or more color pigments (for example, in mammals: red, blue and green).

In Drosophila, the  eye   is composed   of 800   facets  or   ommatidia.  Each
ommatidium contains eight photoreceptor cells (R1-R8):  the R1 to R6 cells are
outer cells,  R7  and R8 inner cells. Each of the three types of cells (R1-R6,
R7 and R8) expresses a specific opsin.

Proteins evolutionary related to opsins include:

 - Squid  retinochrome,  also  known as retinal photoisomerase, which converts
   various isomers of retinal into 11-cis retinal.
 - Mammalian  opsin  3  (Encephalopsin)  that  may  play  a role in encephalic
   photoreception.
 - Mammalian  opsin  4  (Melanopsin)  that may mediate regulation of circadian
   rhythms and acute suppression of pineal melatonin.
 - Mammalian retinal pigment epithelium (RPE) RGR [3], a protein that may also
   act in retinal isomerization.

The attachment  site  for  retinal in the above proteins is a conserved lysine
residue in  the  middle  of  the  seventh  transmembrane helix. The pattern we
developed includes this residue.

-Consensus pattern: [LIVMFWAC]-[PSGAC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-
                    [STACP]-x(2)-[DENF]-[AP]-x(2)-[IY]
                    [K is the retinal binding site]
```

```
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in Swiss-Prot: 2.
-Last update: December 2001 / Pattern and text revised.

[ 1] Applebury M.L., Hargrave P.A.
     Vision Res. 26:1881-1895(1986).
[ 2] Fryxell K.J., Meyerowitz E.M.
     J. Mol. Evol. 33:367-378(1991).
[ 3] Shen D., Jiang M., Hao W., Tao L., Salazar M., Fong H.K.W.
     Biochemistry 33:13117-13125(1994).

+----------------------------------------------------------------------------+
| This PROSITE entry  is copyright  by the Swiss Institute of Bioinformatics |
| (SIB).  There are no restrictions on its use by non-profit institutions as |
| long as its  content  is  in  no  way  modified  and this statement is not |
| removed. Usage by and for commercial entities requires a license agreement |
| (See http://www.isb-sib.ch/announce/ or email to license@isb-sib.ch).      |
+----------------------------------------------------------------------------+
{END}
```

Στην $1^{η}$ γραμμή υπάρχει ο αριθμός πρόσβασης (accession number) του κειμένου και στη $2^{η}$ γραμμή υπάρχει ο αριθμός πρόσβασης (accession number) και ο κωδικός προσδιορισμού (identification code) που χρησιμοποιείται στη SWISS-PROT.

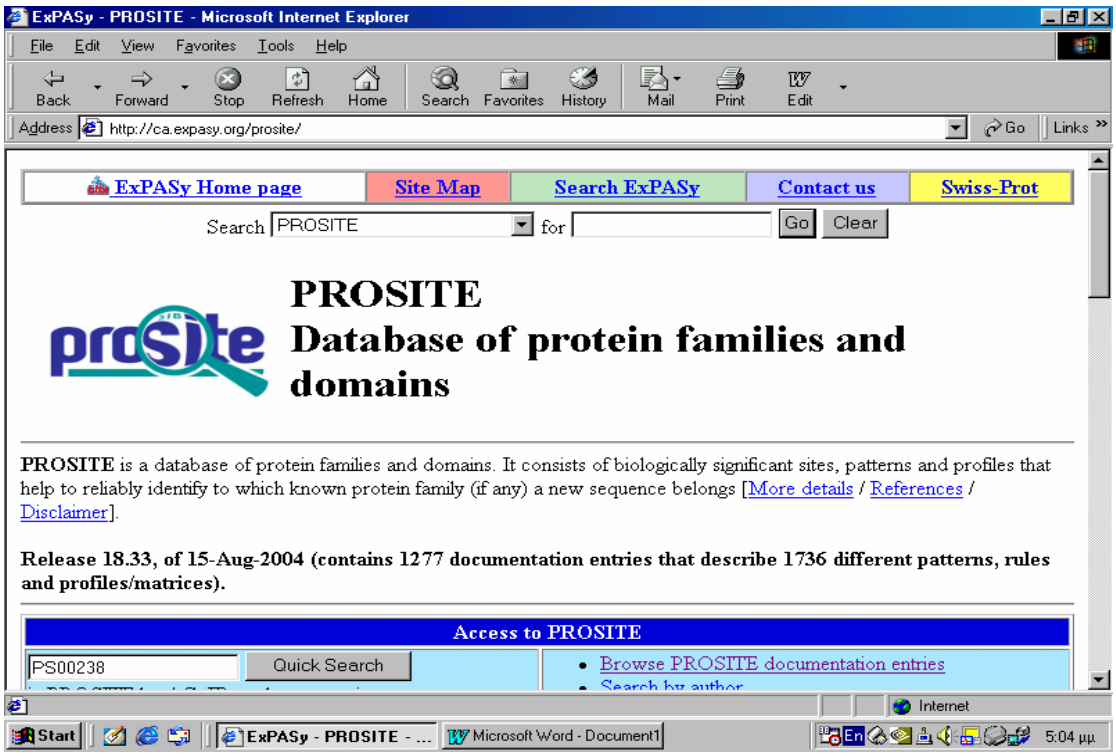
PRINTS

Η PRINTS διατηρείται στο Department of Biochemistry and Molecular Biology του UCL. Η PRINTS στηρίζεται στην αρχή ότι οι περισσότερες οικογένειες πρωτεϊνών χαρακτηρίζονται από πολλαπλά διατηρημένα μοτίβα (multiple reserved motifs), και επομένως πολλές, ή ακόμα και όλες, από αυτές πρέπει να χρησιμοποιηθούν προκειμένου να δημιουργήσουν διαγνωστικές υπογραφές (diagnostic signatures, ή αλλιώς fingerprints) για τα μέλη της κάθε οικογένειας. Στην PRINTS, εάν μία προς αναζήτηση αλληλουχία δεν μπορέσει να ταιριάξει όλα τα μοτίβα σε ένα δεδομένο fingerprint (δακτυλικό αποτύπωμα), το μοντέλο των αντιστοιχήσεων / ταιριασμάτων το οποίο έχει δημιουργηθεί από τα εναπομείναντα μοτίβα επιτρέπει στο χρήστη να κάνει μία διάγνωση.
Σε κάθε εγγραφή της PRINTS, στην αρχή του αρχείου, σε κάθε fingerprint αποδίδεται ένας ID κωδικός, π.χ. το fingerprint (ή αλλιώς signature) για το opsins είναι OPSIN. Κάθε εγγραφή έχει ένα μοναδικό AC, το οποίο είναι της μορφής PR00000 (μπορεί να χρησιμοποιηθεί ένα AC της PROSITE) , καθώς και μία ένδειξη του αριθμού των μοτίβων στο fingerprint (στην περίπτωση του OPSIN είναι 3).
Υπάρχει μία ενότητα με πληροφορίες σχετικά με τη διαγνωστική επίδοση του fingerprint και των μοτίβων που το αποτελούν (για το PS00238, 123 αλληλουχίες ταίριαξαν και τα 3 στοιχεία του fingerprint, ενώ 7 αλληλουχίες ταίριαξαν μόνο 2 μοτίβα).
Η επόμενη ενότητα παρουσιάζει όλες τις πρωτεΐνες οι οποίες προσδιορίζονται χρησιμοποιώντας fingerprint όταν το PS00238 αντιστοιχίζεται πολλαπλά στη βάση δεδομένων.
Στις επόμενες ενότητες παρέχονται τα μοτίβα που χρησιμοποιούνται σαν ‘αρχή’ (seed) στην επαναλαμβανόμενη αναζήτηση στη βάση δεδομένων.
Στην τελευταία ενότητα (FINAL MOTIF SETS) παρέχονται τα τρία μοτίβα για την κάθε σχετική πρωτεΐνη, μαζί με την τοποθεσία στην αρχική αλληλουχία (parent sequence-ST), καθώς και τον αριθμό των residues από τον προηγούμενο γείτονα (INT).

ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

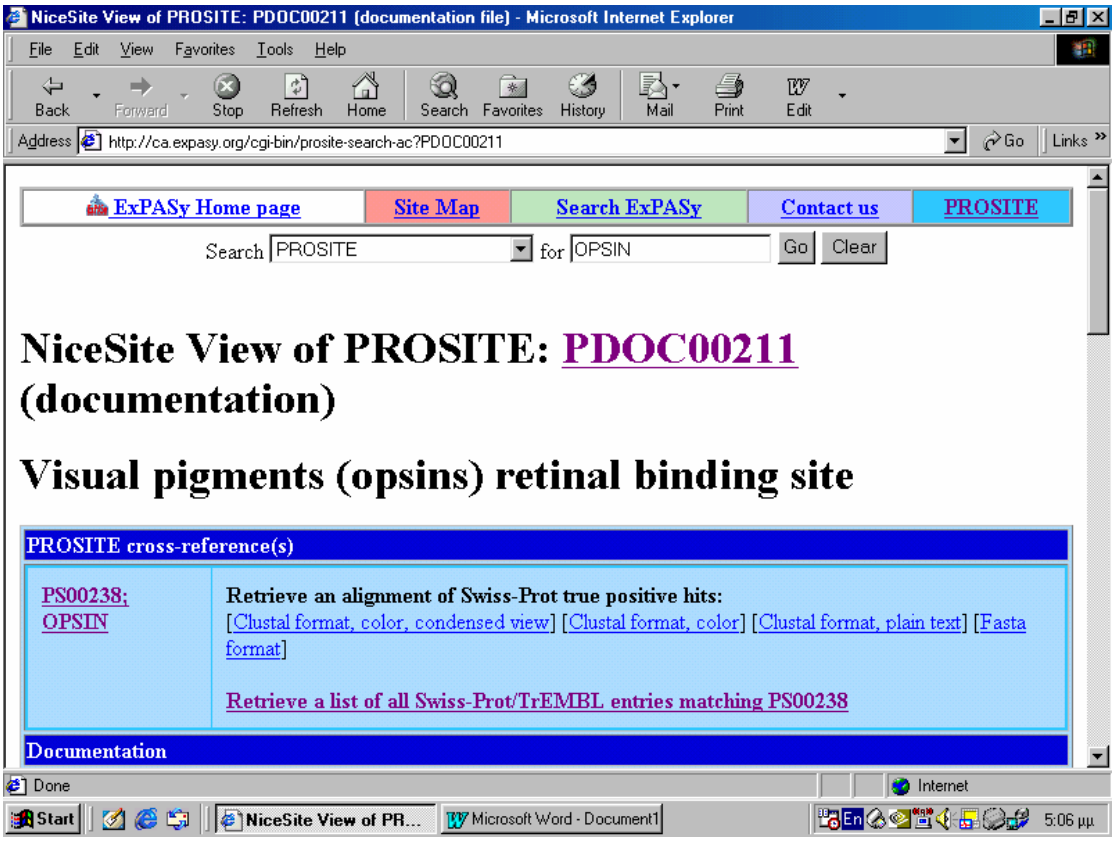
Δευτεροταγείς βάσεις δεδομένων – Ανάκτηση από την PROSITE

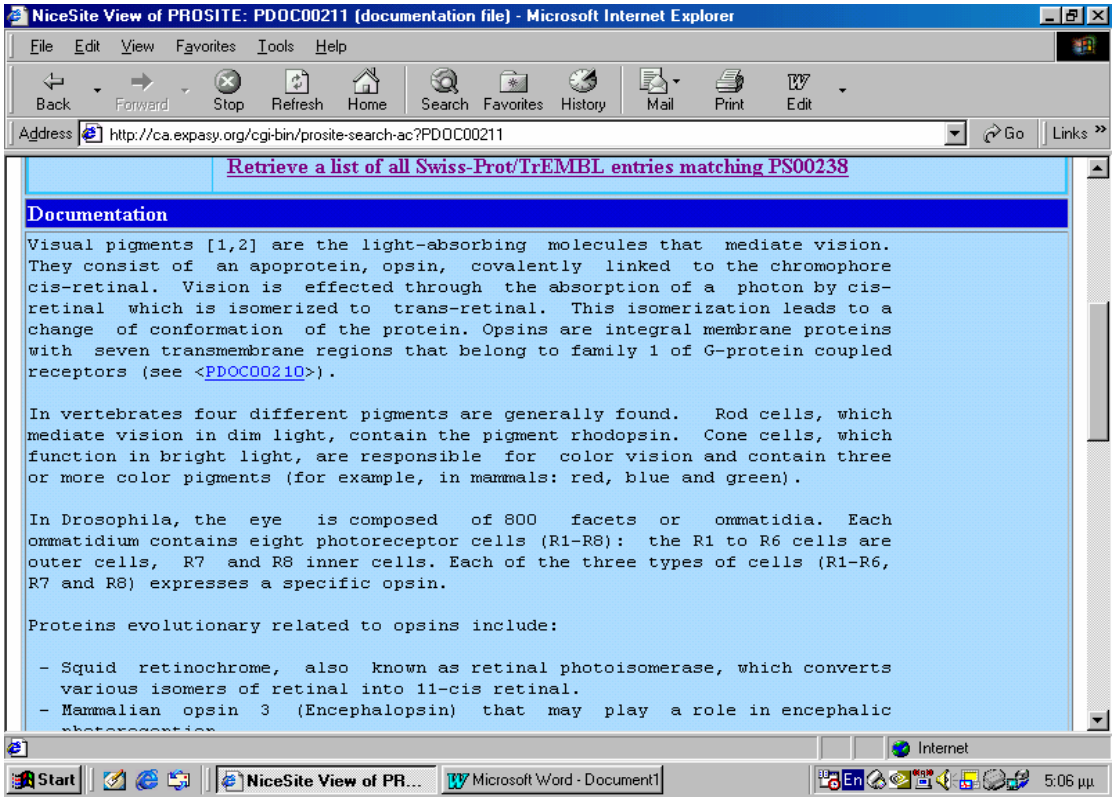
Στο Internet Explorer πληκτρολογείστε την διεύθυνση της PROSITE: au.expasy.org/prosite/. Για να αναγνωρίσετε τις πρωτεΐνες που σχετίζονται με την οικογένεια πρωτεϊνών OPSIN βασιζόμενη σε μία regular expression ως διαγνωστικό

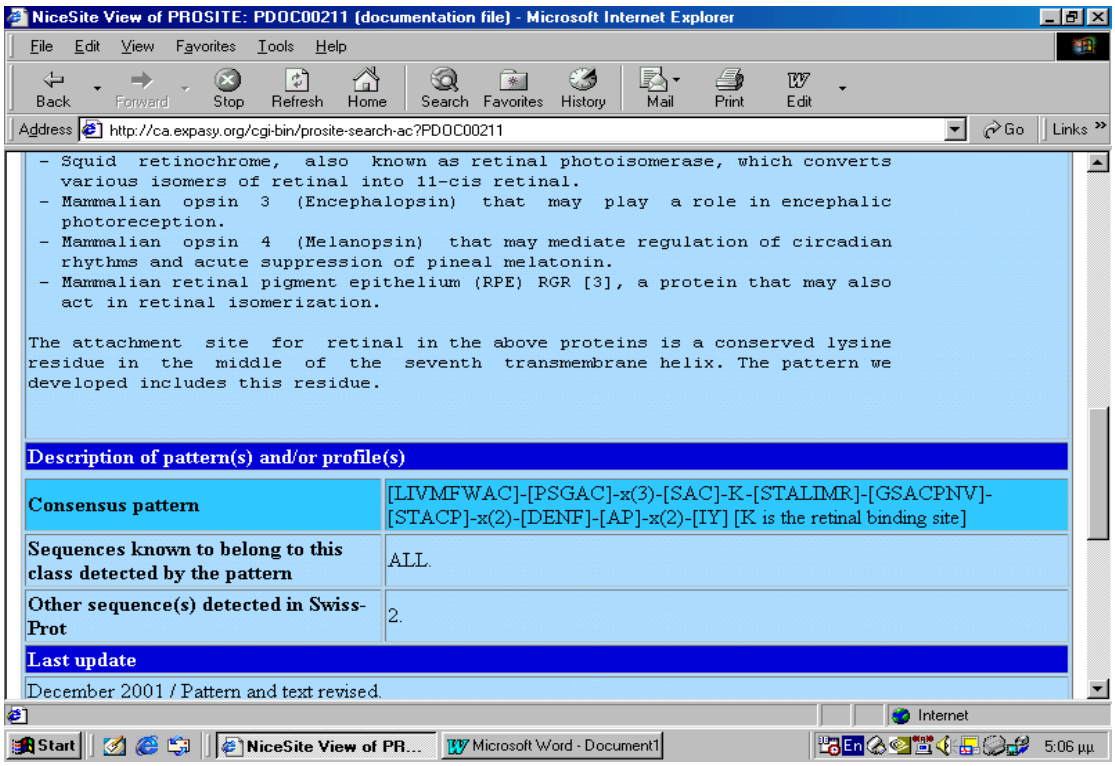


εργαλείο, πληκτρολογήστε τον accession code PS00238 στο πλαίσιο παρακάτω και επιλέξτε “Quick Search”.

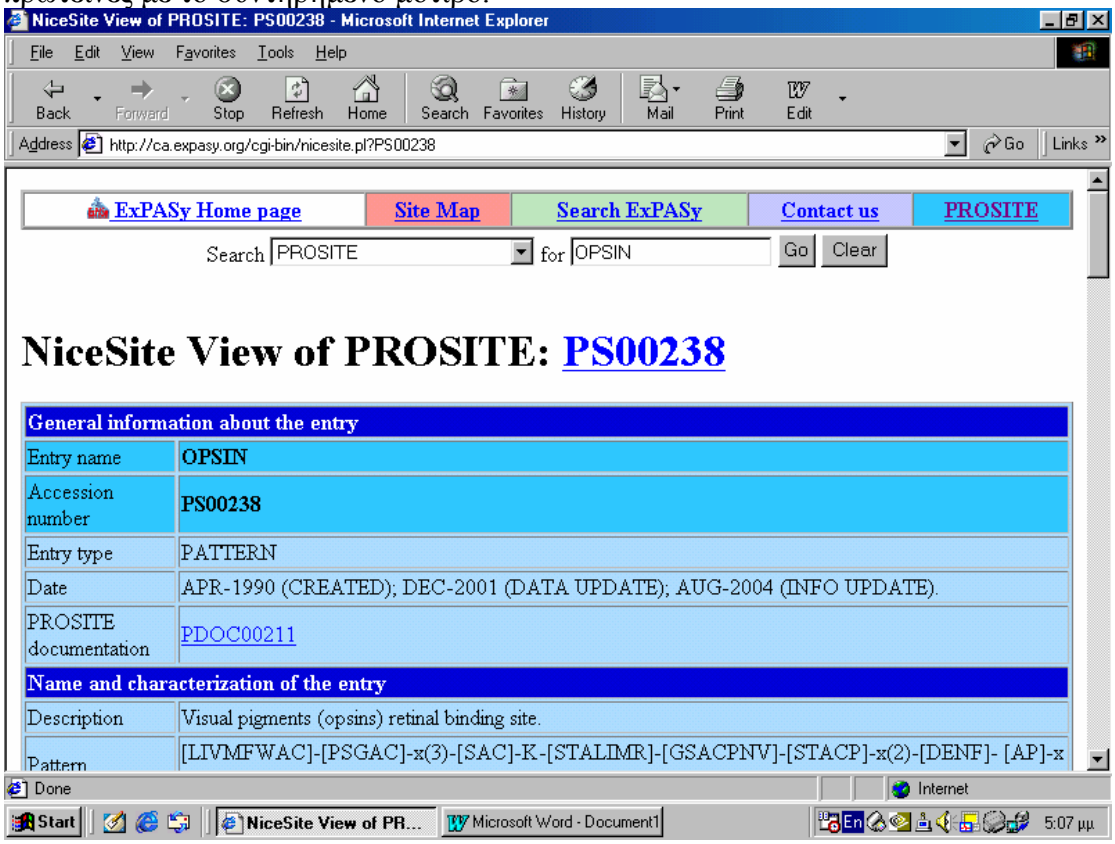
Τότε η ιστοσελίδα με όλες τις πληροφορίες για την OPSIN προβάλλεται, μαζί με τη σύντομη περιγραφή της (consensus pattern).

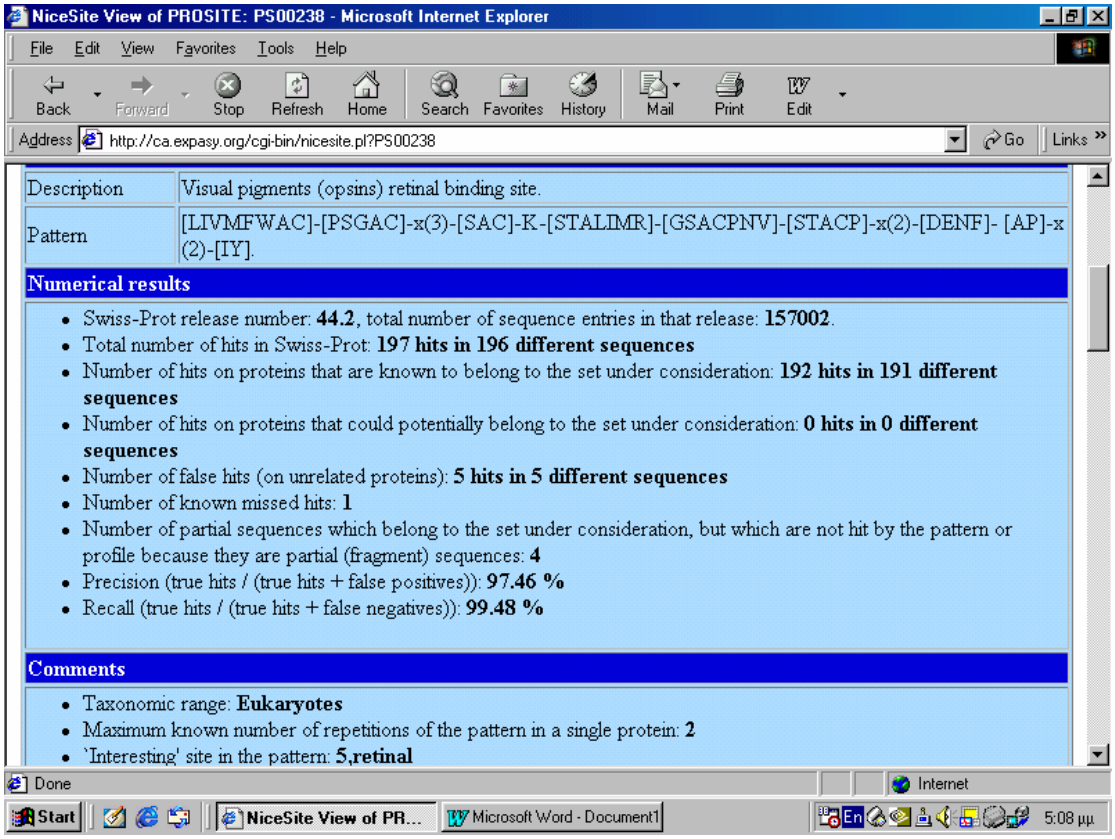


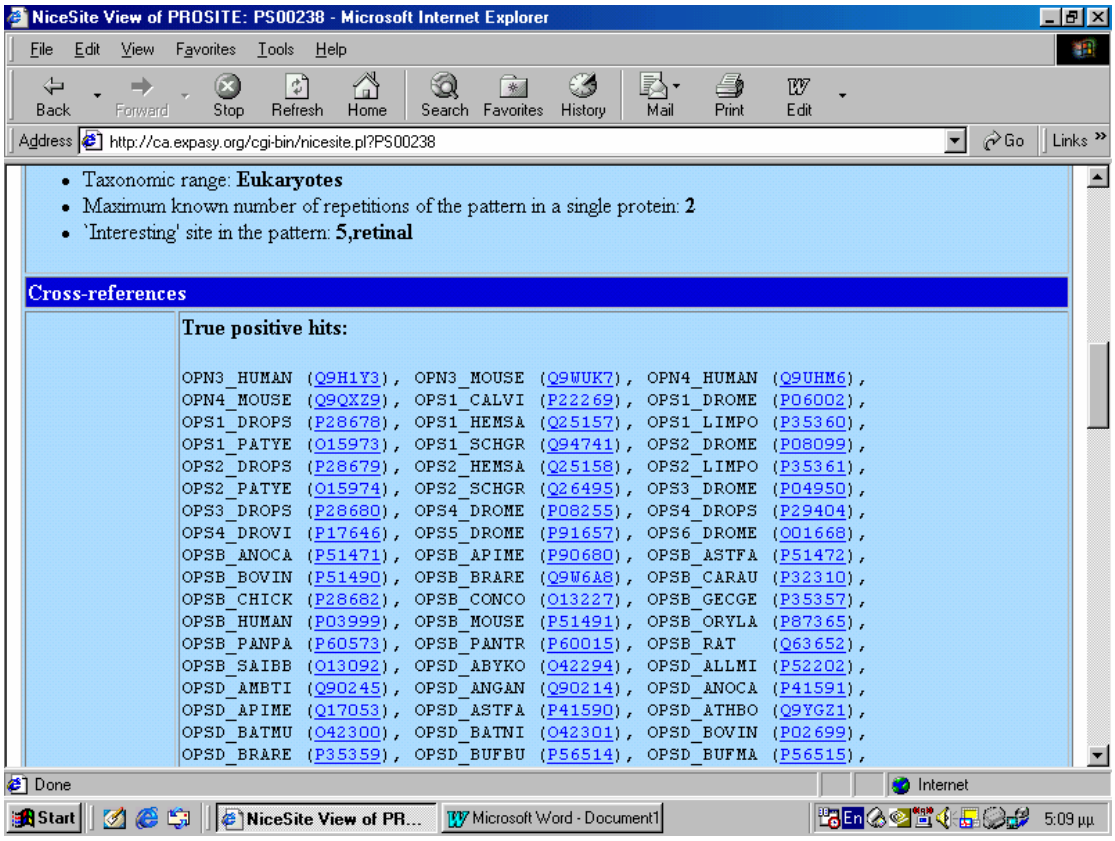

NiceSite View of PROSITE: PDOC00211 (documentation file) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/prosite-search-ac?PDOC00211
Go Links
Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS00238
Documentation
Visual pigments [1,2] are the light-absorbing molecules that mediate vision. They consist of an apoprotein, opsin, covalently linked to the chromophore cis-retinal. Vision is effected through the absorption of a photon by cis-retinal which is isomerized to trans-retinal. This isomerization leads to a change of conformation of the protein. Opsins are integral membrane proteins with seven transmembrane regions that belong to family 1 of G-protein coupled receptors (see <PDOC00210>).
In vertebrates four different pigments are generally found. Rod cells, which mediate vision in dim light, contain the pigment rhodopsin. Cone cells, which function in bright light, are responsible for color vision and contain three or more color pigments (for example, in mammals: red, blue and green).
In Drosophila, the eye is composed of 800 facets or ommatidia. Each ommatidium contains eight photoreceptor cells (R1-R8): the R1 to R6 cells are outer cells, R7 and R8 inner cells. Each of the three types of cells (R1-R6, R7 and R8) expresses a specific opsin.
Proteins evolutionary related to opsins include:
- Squid retinochrome, also known as retinal photoisomerase, which converts various isomers of retinal into 11-cis retinal.
- Mammalian opsin 3 (Encephalopsin) that may play a role in encephalic
Internet
Start
NiceSite View of PR...
Microsoft Word - Document1
5:06 µµ

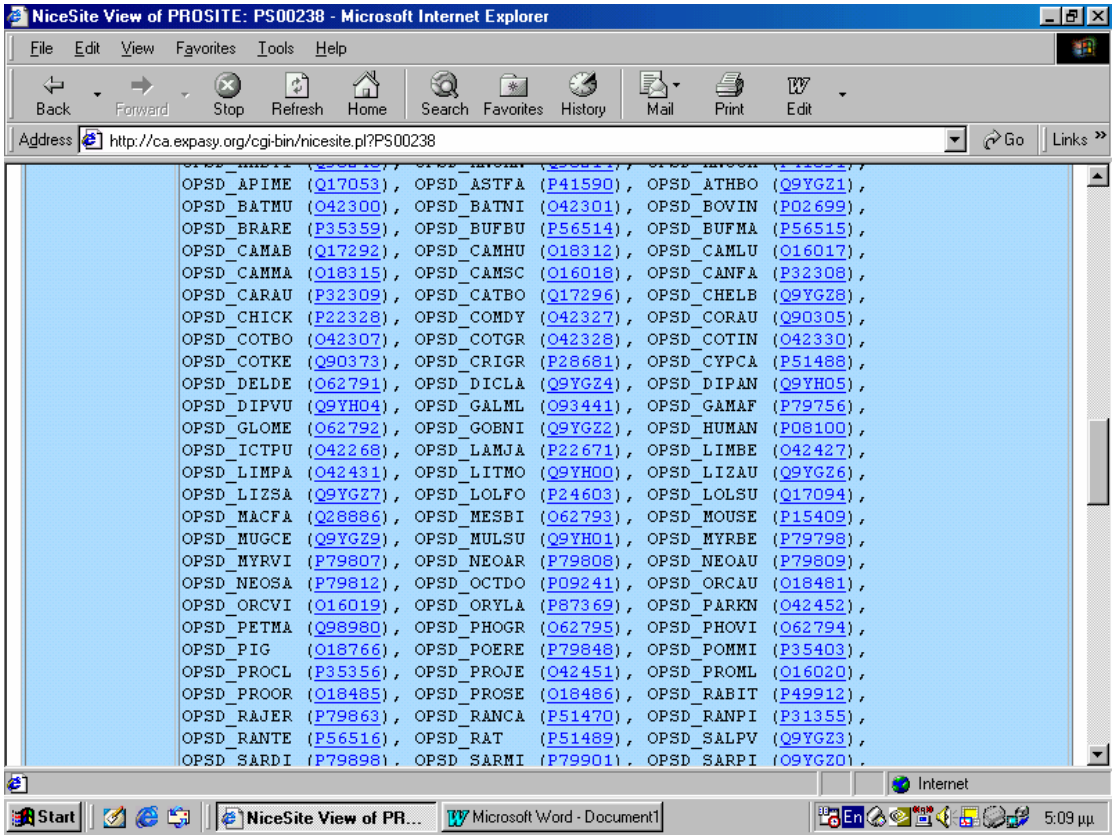

NiceSite View of PROSITE: PDOC00211 (documentation file) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/prosite-search-ac?PDOC00211
Go Links
- Squid retinochrome, also known as retinal photoisomerase, which converts various isomers of retinal into 11-cis retinal.
- Mammalian opsin 3 (Encephalopsin) that may play a role in encephalic photoreception.
- Mammalian opsin 4 (Melanopsin) that may mediate regulation of circadian rhythms and acute suppression of pineal melatonin.
- Mammalian retinal pigment epithelium (RPE) RGR [3], a protein that may also act in retinal isomerization.
The attachment site for retinal in the above proteins is a conserved lysine residue in the middle of the seventh transmembrane helix. The pattern we developed includes this residue.
Description of pattern(s) and/or profile(s)
Consensus pattern
[LIVMFWAC]-[PSGAC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-[STACP]-x(2)-[DENF]-[AP]-x(2)-[IY] [K is the retinal binding site]
Sequences known to belong to this class detected by the pattern
ALL.
Other sequence(s) detected in Swiss-Prot
2.
Last update
December 2001 / Pattern and text revised.
Internet
Start
NiceSite View of PR...
Microsoft Word - Document1
5:06 µµ

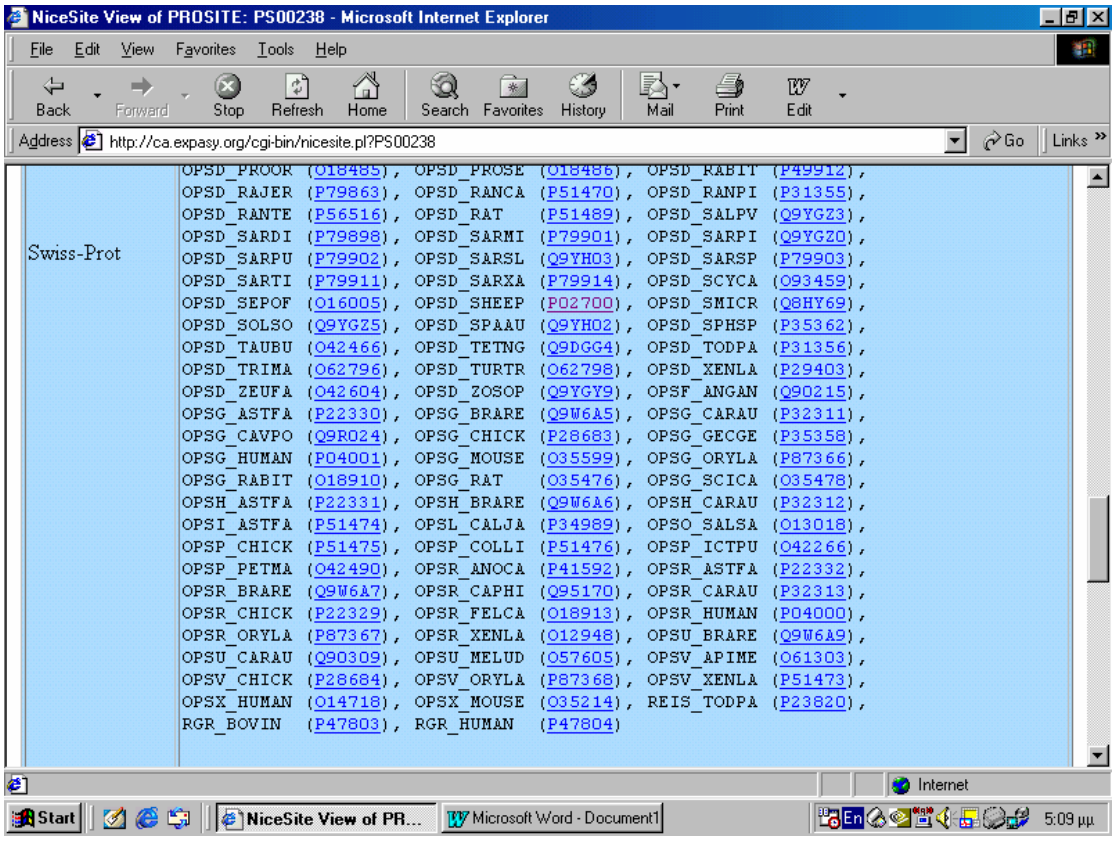
Πατώντας AC PS00238 στην παραπάνω σελίδα, παρουσιάζονται τα αποτελέσματα της αναζήτησης στην PROSITE, τα οποία περιλαμβάνουν όλες τις σχετιζόμενες πρωτεΐνες με το συντηρημένο μοτίβο.

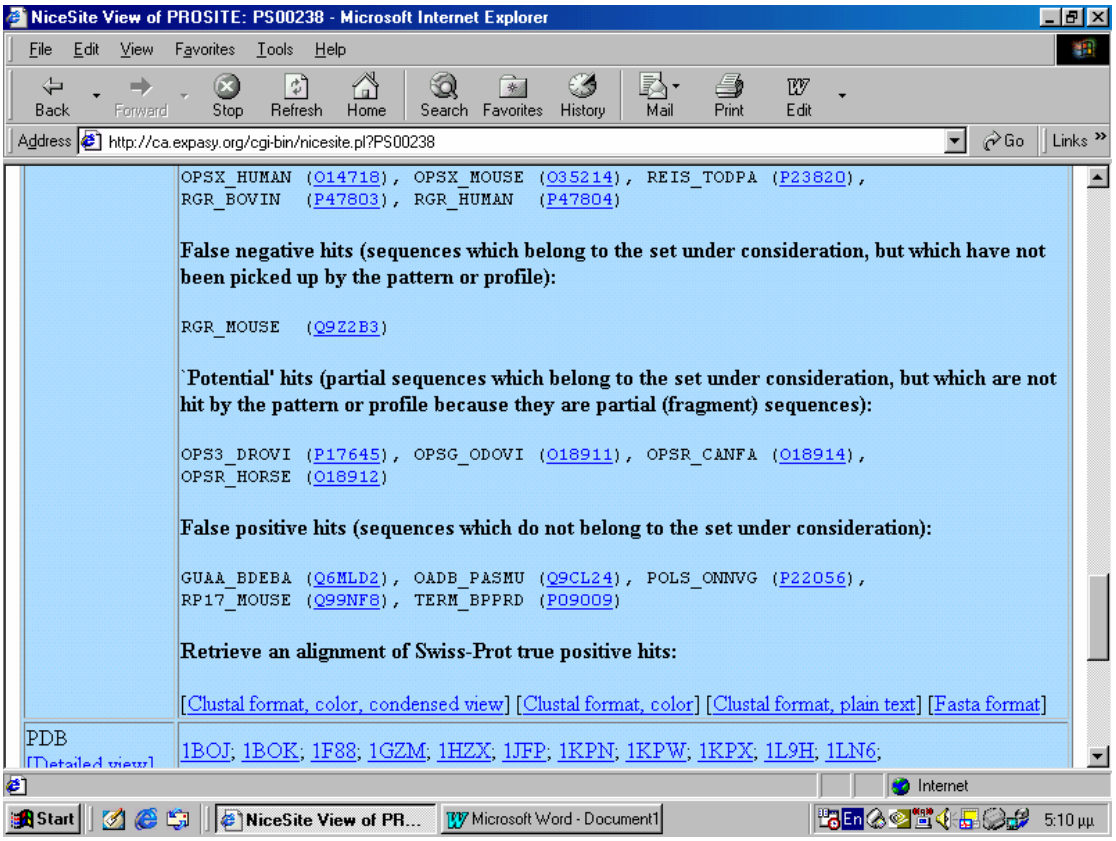


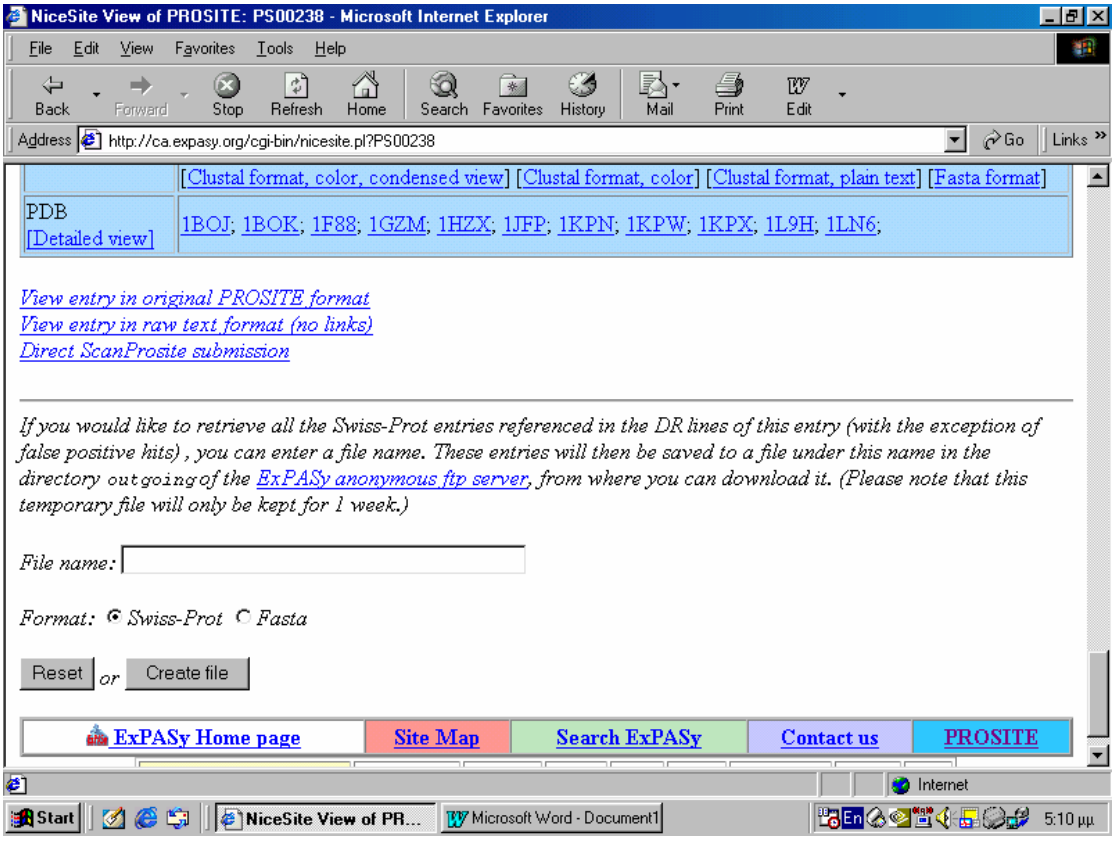

NiceSite View of PROSITE: PS00238 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/nicesite.pl?PS00238
Go Links
Description Visual pigments (opsins) retinal binding site.
Pattern [LIVMFWAC]-[PSGAC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-[STACP]-x(2)-[DENF]- [AP]-x(2)-[IY].
Numerical results
Swiss-Prot release number: 44.2, total number of sequence entries in that release: 157002.
Total number of hits in Swiss-Prot: 197 hits in 196 different sequences
Number of hits on proteins that are known to belong to the set under consideration: 192 hits in 191 different sequences
Number of hits on proteins that could potentially belong to the set under consideration: 0 hits in 0 different sequences
Number of false hits (on unrelated proteins): 5 hits in 5 different sequences
Number of known missed hits: 1
Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: 4
Precision (true hits / (true hits + false positives)): 97.46 %
Recall (true hits / (true hits + false negatives)): 99.48 %
Comments
Taxonomic range: Eukaryotes
Maximum known number of repetitions of the pattern in a single protein: 2
'Interesting' site in the pattern: 5,retinal
Done
Internet
Start
NiceSite View of PR...
Microsoft Word - Document1
5:08 μμ


NiceSite View of PROSITE: PS00238 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/nicesite.pl?PS00238
Go Links
Taxonomic range: Eukaryotes
Maximum known number of repetitions of the pattern in a single protein: 2
'Interesting' site in the pattern: 5,retinal
Cross-references
True positive hits:
OPN3_HUMAN (Q9H1Y3), OPN3_MOUSE (Q9WUK7), OPN4_HUMAN (Q9UHM6),
OPN4_MOUSE (Q9QXZ9), OPS1_CALVI (P22269), OPS1_DROME (P06002),
OPS1_DROPS (P28678), OPS1_HEMSA (Q25157), OPS1_LIMPO (P35360),
OPS1_PATYE (O15973), OPS1_SCHGR (Q94741), OPS2_DROME (P08099),
OPS2_DROPS (P28679), OPS2_HEMSA (Q25158), OPS2_LIMPO (P35361),
OPS2_PATYE (O15974), OPS2_SCHGR (Q26495), OPS3_DROME (P04950),
OPS3_DROPS (P28680), OPS4_DROME (P08255), OPS4_DROPS (P29404),
OPS4_DROVI (P17646), OPS5_DROME (P91657), OPS6_DROME (O01668),
OPSB_ANOCA (P51471), OPSB_APIME (P90680), OPSB_ASTFA (P51472),
OPSB_BOVIN (P51490), OPSB_BRARE (Q9W6A8), OPSB_CARAU (P32310),
OPSB_CHICK (P28682), OPSB_CONCO (O13227), OPSB_GECGE (P35357),
OPSB_HUMAN (P03999), OPSB_MOUSE (P51491), OPSB_ORYLA (P87365),
OPSB_PANPA (P60573), OPSB_PANTR (P60015), OPSB_RAT (Q63652),
OPSB_SAIBB (O13092), OPSD_ABYKO (O42294), OPSD_ALLMI (P52202),
OPSD_AMBTI (Q90245), OPSD_ANGAN (Q90214), OPSD_ANOCA (P41591),
OPSD_APIME (Q17053), OPSD_ASTFA (P41590), OPSD_ATHBO (Q9YGZ1),
OPSD_BATMU (O42300), OPSD_BATNI (O42301), OPSD_BOVIN (P02699),
OPSD_BRARE (P35359), OPSD_BUFBU (P56514), OPSD_BUFMA (P56515),
Done
Internet
Start
NiceSite View of PR...
Microsoft Word - Document1
5:09 μμ

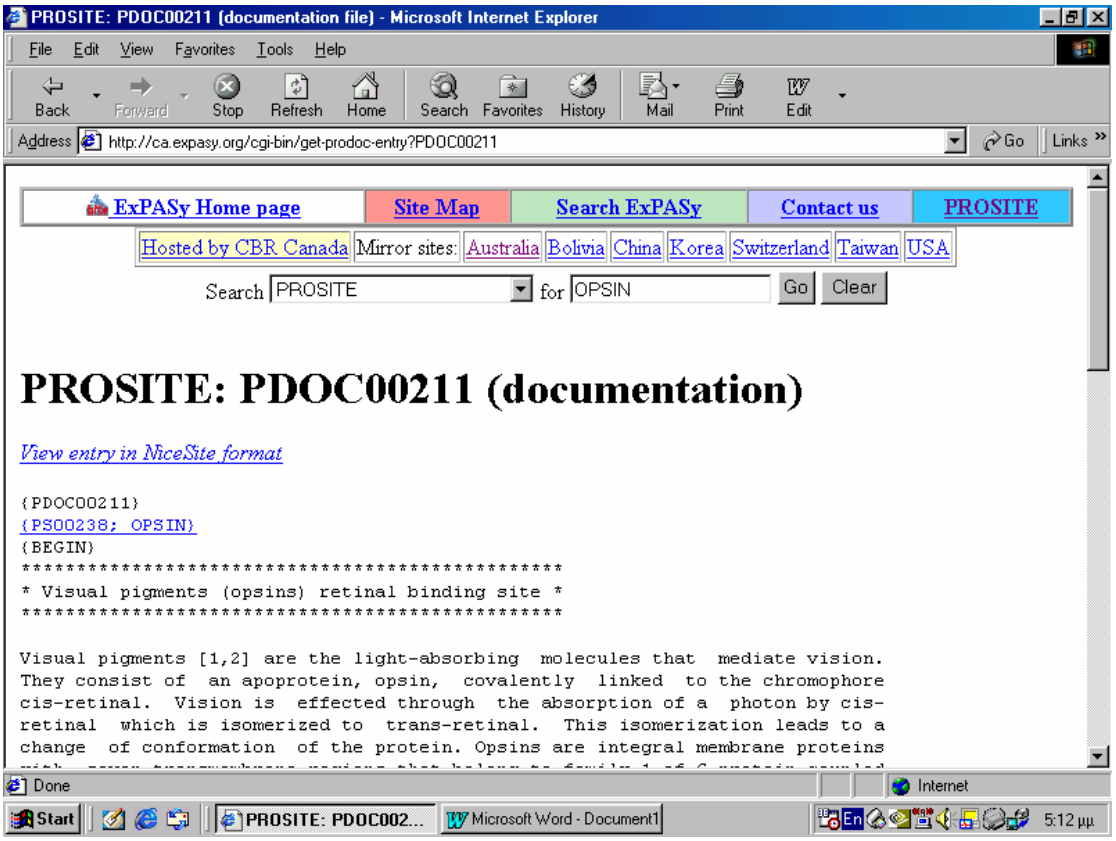

NiceSite View of PROSITE: PS00238 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/nicesite.pl?PS00238
Go Links
OPSD_APIME (Q17053), OPSD_ASTFA (P41590), OPSD_ATHBO (Q9YGZ1),
OPSD_BATMU (O42300), OPSD_BATNI (O42301), OPSD_BOVIN (P02699),
OPSD_BRARE (P35359), OPSD_BUFBU (P56514), OPSD_BUFMA (P56515),
OPSD_CAMAB (Q17292), OPSD_CAMHU (O18312), OPSD_CAMLU (O16017),
OPSD_CAMMA (O18315), OPSD_CAMSC (O16018), OPSD_CANFA (P32308),
OPSD_CARAU (P32309), OPSD_CATBO (Q17296), OPSD_CHELB (Q9YGZ8),
OPSD_CHICK (P22328), OPSD_COMDY (O42327), OPSD_CORAU (Q90305),
OPSD_COTBO (O42307), OPSD_COTGR (O42328), OPSD_COTIN (O42330),
OPSD_COTKE (Q90373), OPSD_CRIGR (P28681), OPSD_CYPCA (P51488),
OPSD_DELDE (O62791), OPSD_DICLA (Q9YGZ4), OPSD_DIPAN (Q9YH05),
OPSD_DIPVU (Q9YH04), OPSD_GALML (O93441), OPSD_GAMAF (P79756),
OPSD_GLOME (O62792), OPSD_GOBNI (Q9YGZ2), OPSD_HUMAN (P08100),
OPSD_ICTPU (O42268), OPSD_LAMJA (P22671), OPSD_LIMBE (O42427),
OPSD_LIMPA (O42431), OPSD_LITMO (Q9YH00), OPSD_LIZAU (Q9YGZ6),
OPSD_LIZSA (Q9YGZ7), OPSD_LOLFO (P24603), OPSD_LOLSU (Q17094),
OPSD_MACFA (Q28886), OPSD_MESBI (O62793), OPSD_MOUSE (P15409),
OPSD_MUGCE (Q9YGZ9), OPSD_MULSU (Q9YH01), OPSD_MYRBE (P79798),
OPSD_MYRVI (P79807), OPSD_NEOAR (P79808), OPSD_NEOAU (P79809),
OPSD_NEOSA (P79812), OPSD_OCTDO (P09241), OPSD_ORCAU (O18481),
OPSD_ORCVI (O16019), OPSD_ORYLA (P87369), OPSD_PARKN (O42452),
OPSD_PETMA (Q98980), OPSD_PHOGR (O62795), OPSD_PHOVI (O62794),
OPSD_PIG (O18766), OPSD_POERE (P79848), OPSD_POMMI (P35403),
OPSD_PROCL (P35356), OPSD_PROJE (O42451), OPSD_PROML (O16020),
OPSD_PROOR (O18485), OPSD_PROSE (O18486), OPSD_RABIT (P49912),
OPSD_RAJER (P79863), OPSD_RANCA (P51470), OPSD_RANPI (P31355),
OPSD_RANTE (P56516), OPSD_RAT (P51489), OPSD_SALPV (Q9YGZ3),
OPSD_SARDI (P79898), OPSD_SARMI (P79901), OPSD_SARPI (Q9YGZ0),
Internet
Start
NiceSite View of PR...
Microsoft Word - Document1
5:09 μμ

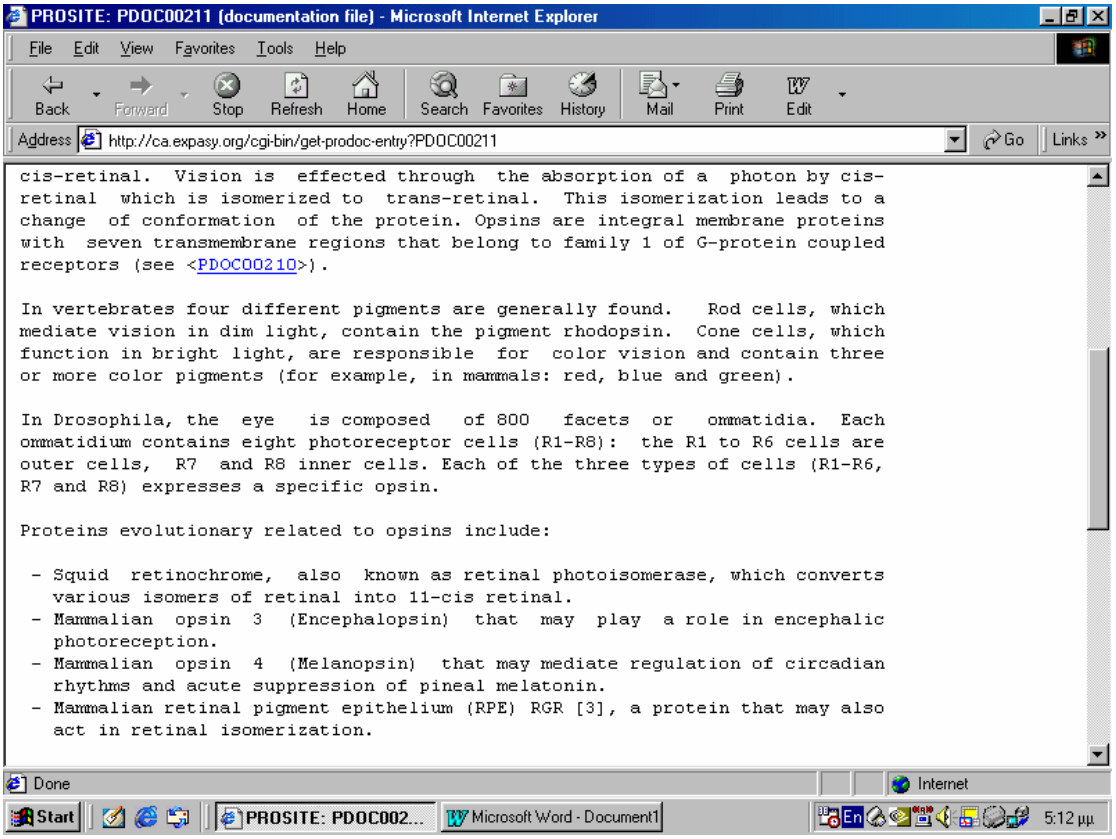

NiceSite View of PROSITE: PS00238 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/nicesite.pl?PS00238
Go Links
OPSD_PROOR (O18485), OPSD_PROSE (O18486), OPSD_RABIT (P49912),
OPSD_RAJER (P79863), OPSD_RANCA (P51470), OPSD_RANPI (P31355),
OPSD_RANTE (P56516), OPSD_RAT (P51489), OPSD_SALPV (Q9YGZ3),
OPSD_SARDI (P79898), OPSD_SARMI (P79901), OPSD_SARPI (Q9YGZ0),
Swiss-Prot
OPSD_SARPU (P79902), OPSD_SARSL (Q9YH03), OPSD_SARSP (P79903),
OPSD_SARTI (P79911), OPSD_SARXA (P79914), OPSD_SCYCA (O93459),
OPSD_SEPOF (O16005), OPSD_SHEEP (P02700), OPSD_SMICR (Q8HY69),
OPSD_SOLSO (Q9YGZ5), OPSD_SPAAU (Q9YH02), OPSD_SPHSP (P35362),
OPSD_TAUBU (O42466), OPSD_TETNG (Q9DGG4), OPSD_TODPA (P31356),
OPSD_TRIMA (O62796), OPSD_TURTR (O62798), OPSD_XENLA (P29403),
OPSD_ZEUFA (O42604), OPSD_ZOSOP (Q9YGY9), OPSF_ANGAN (Q90215),
OPSG_ASTFA (P22330), OPSG_BRARE (Q9W6A5), OPSG_CARAU (P32311),
OPSG_CAVPO (Q9R024), OPSG_CHICK (P28683), OPSG_GECGE (P35358),
OPSG_HUMAN (P04001), OPSG_MOUSE (O35599), OPSG_ORYLA (P87366),
OPSG_RABIT (O18910), OPSG_RAT (O35476), OPSG_SCICA (O35478),
OPSH_ASTFA (P22331), OPSH_BRARE (Q9W6A6), OPSH_CARAU (P32312),
OPSI_ASTFA (P51474), OPSL_CALJA (P34989), OPSO_SALSA (O13018),
OPSP_CHICK (P51475), OPSP_COLLI (P51476), OPSP_ICTPU (O42266),
OPSP_PETMA (O42490), OPSR_ANOCA (P41592), OPSR_ASTFA (P22332),
OPSR_BRARE (Q9W6A7), OPSR_CAPHI (Q95170), OPSR_CARAU (P32313),
OPSR_CHICK (P22329), OPSR_FELCA (O18913), OPSR_HUMAN (P04000),
OPSR_ORYLA (P87367), OPSR_XENLA (O12948), OPSU_BRARE (Q9W6A9),
OPSU_CARAU (Q90309), OPSU_MELUD (O57605), OPSV_APIME (O61303),
OPSV_CHICK (P28684), OPSV_ORYLA (P87368), OPSV_XENLA (P51473),
OPSX_HUMAN (O14718), OPSX_MOUSE (O35214), REIS_TODPA (P23820),
RGR_BOVIN (P47803), RGR_HUMAN (P47804)
Internet
Start
NiceSite View of PR...
Microsoft Word - Document1
5:09 μμ

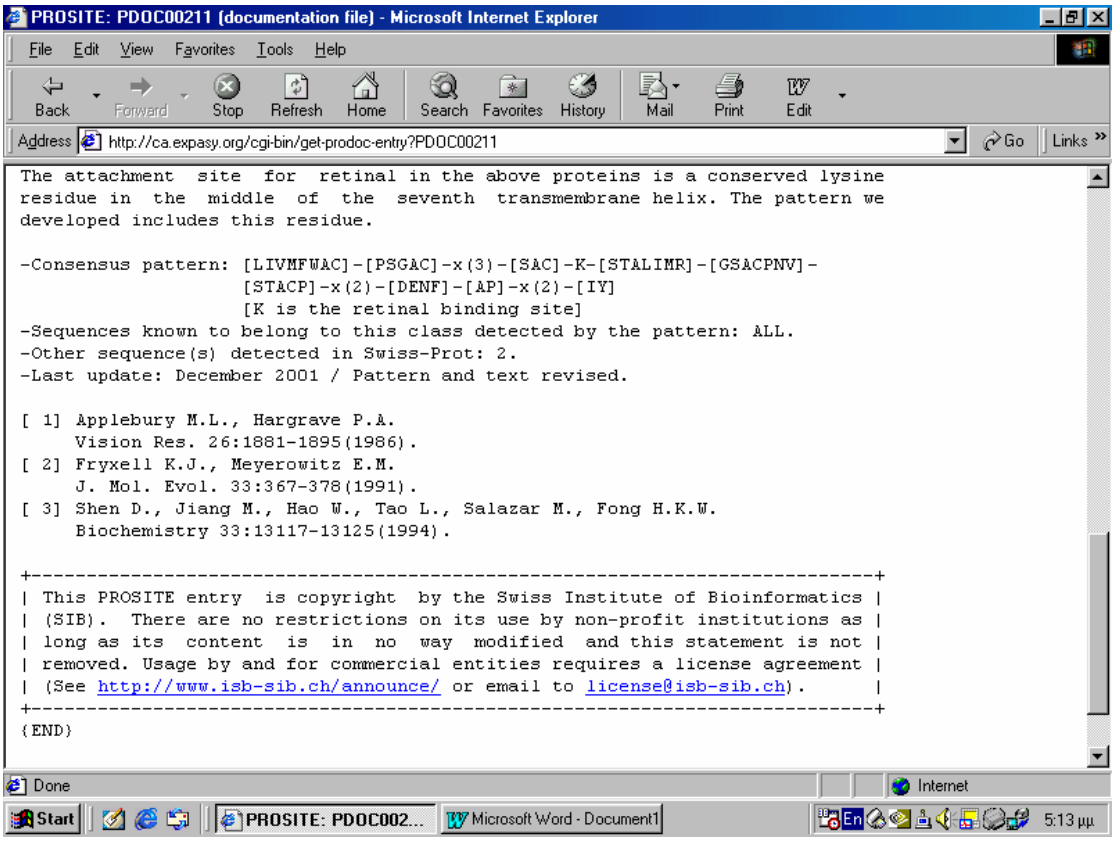

NiceSite View of PROSITE: PS00238 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/nicesite.pl?PS00238
Go Links
OPSX_HUMAN (O14718), OPSX_MOUSE (O35214), REIS_TODPA (P23820),
RGR_BOVIN (P47803), RGR_HUMAN (P47804)
False negative hits (sequences which belong to the set under consideration, but which have not been picked up by the pattern or profile):
RGR_MOUSE (Q9Z2B3)
`Potential' hits (partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences):
OPS3_DROVI (P17645), OPSG_ODOVI (O18911), OPSR_CANFA (O18914),
OPSR_HORSE (O18912)
False positive hits (sequences which do not belong to the set under consideration):
GUAA_BDEBA (Q6MLD2), OADB_PASMU (Q9CL24), POLS_ONNVG (P22056),
RP17_MOUSE (Q99NF8), TERM_BPPRD (P09009)
Retrieve an alignment of Swiss-Prot true positive hits:
[Clustal format, color, condensed view] [Clustal format, color] [Clustal format, plain text] [Fasta format]
PDB
1BOJ; 1BOK; 1F88; 1GZM; 1HZX; 1JFP; 1KPN; 1KPW; 1KPX; 1L9H; 1LN6;
Internet
Start
NiceSite View of PR...
Microsoft Word - Document1
5:10 μμ


NiceSite View of PROSITE: PS00238 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/nicesite.pl?PS00238
Go Links
[Clustal format, color, condensed view] [Clustal format, color] [Clustal format, plain text] [Fasta format]
PDB [Detailed view]
1BOJ; 1BOK; 1F88; 1GZM; 1HZX; 1JFP; 1KPN; 1KPW; 1KPX; 1L9H; 1LN6;
View entry in original PROSITE format
View entry in raw text format (no links)
Direct ScanProsite submission
If you would like to retrieve all the Swiss-Prot entries referenced in the DR lines of this entry (with the exception of false positive hits) , you can enter a file name. These entries will then be saved to a file under this name in the directory outgoing of the ExPASy anonymous ftp server, from where you can download it. (Please note that this temporary file will only be kept for 1 week.)
File name:
Format: Swiss-Prot Fasta
Reset or Create file
ExPASy Home page
Site Map
Search ExPASy
Contact us
PROSITE
Internet
Start
NiceSite View of PR...
Microsoft Word - Document1
5:10 µµ

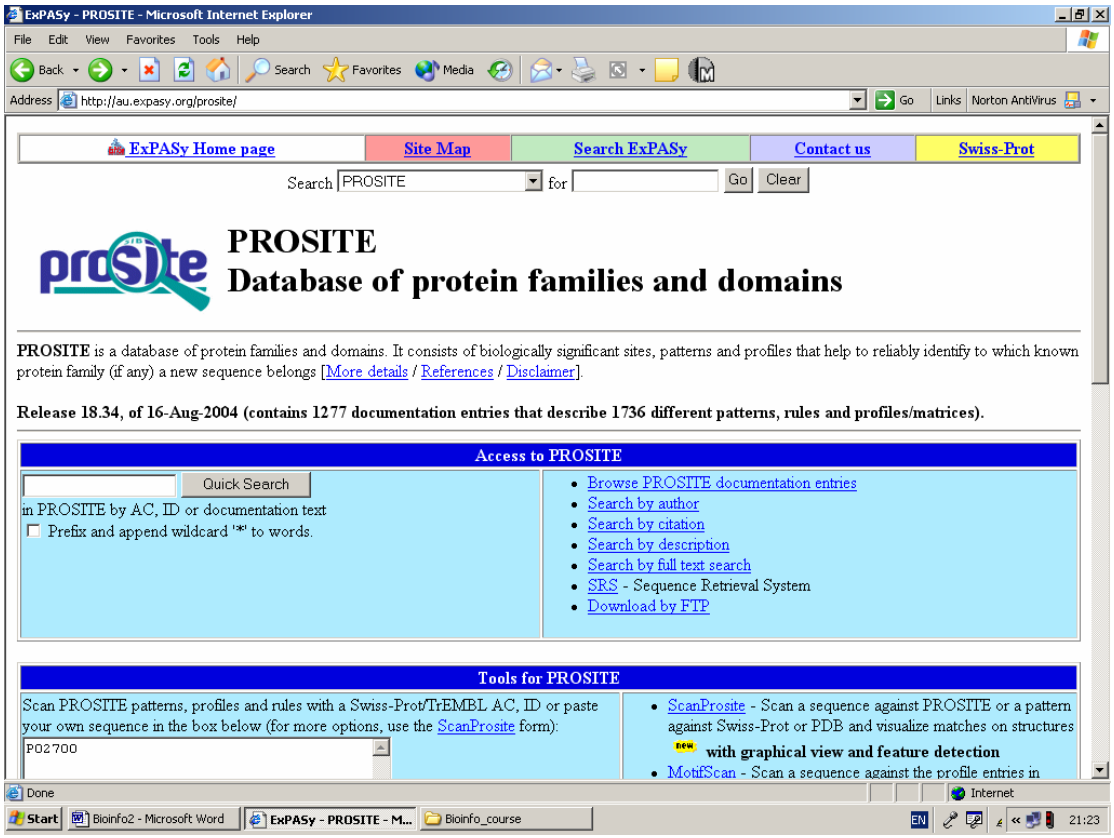
Επιλέγοντας PDOC00211 στην παραπάνω σελίδα, παρουσιάζεται σε ελεύθερη μορφή κειμένου το αρχείο για την OPSIN.

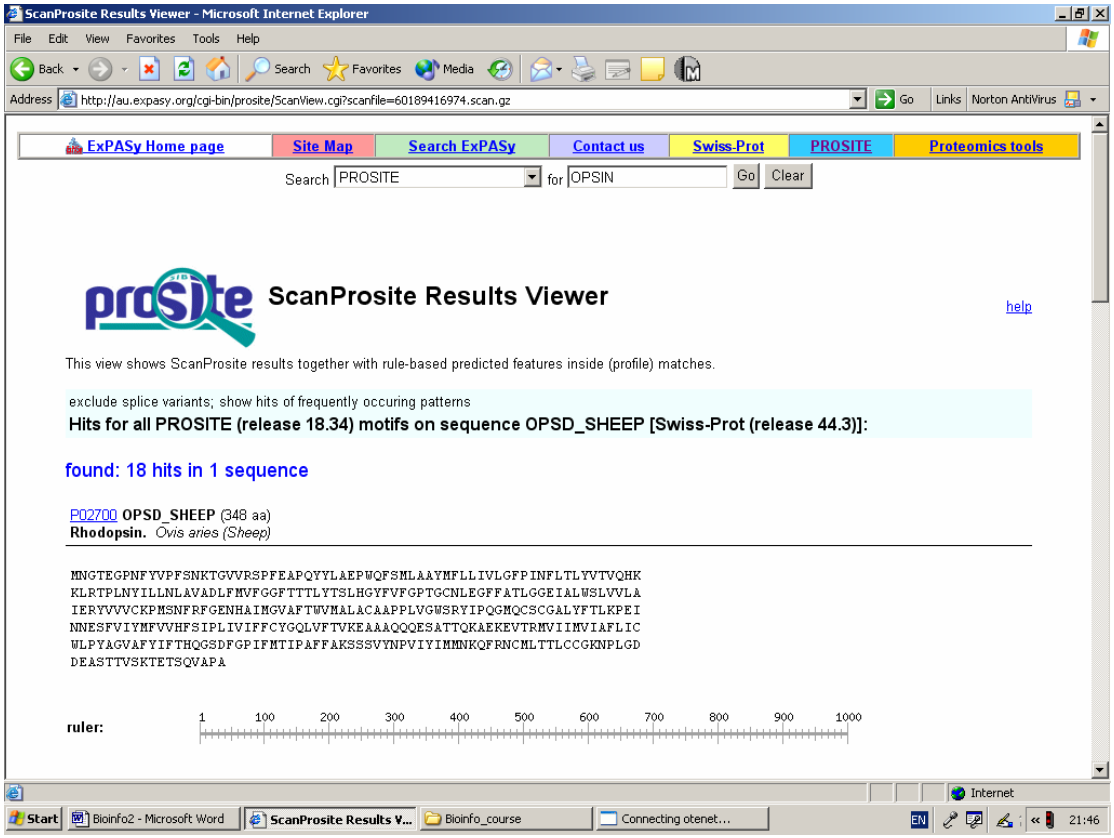


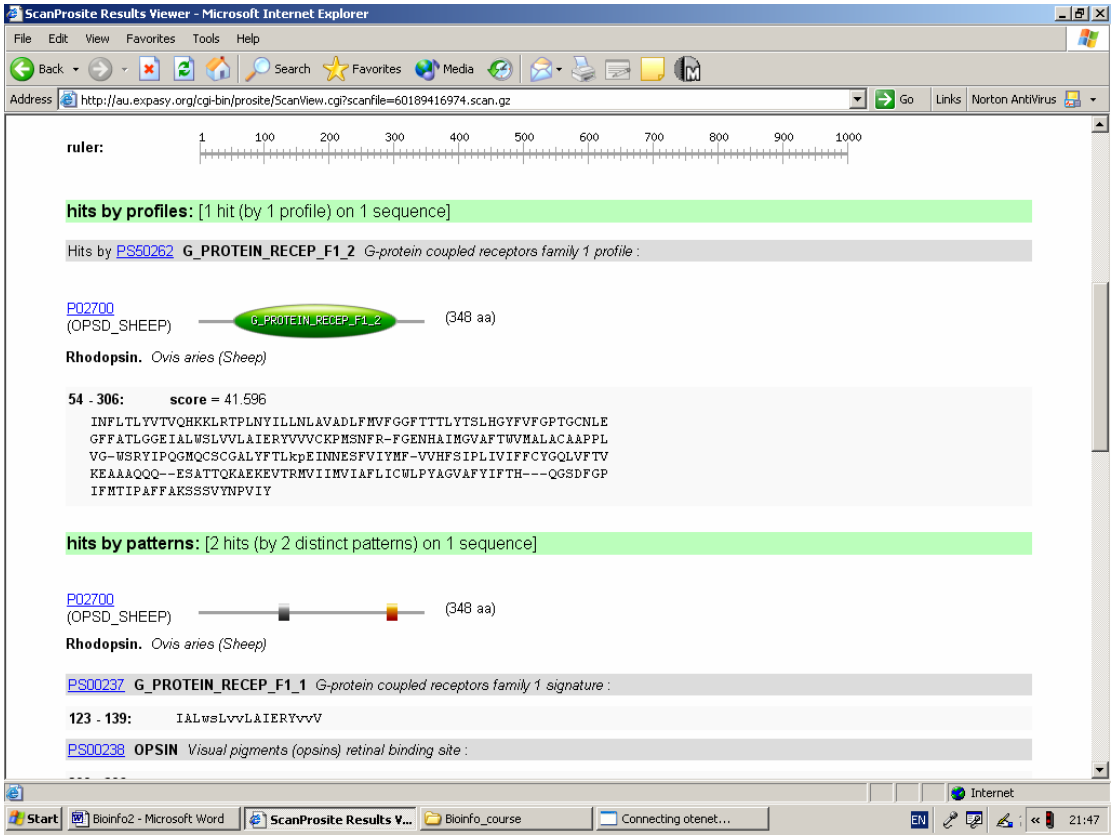


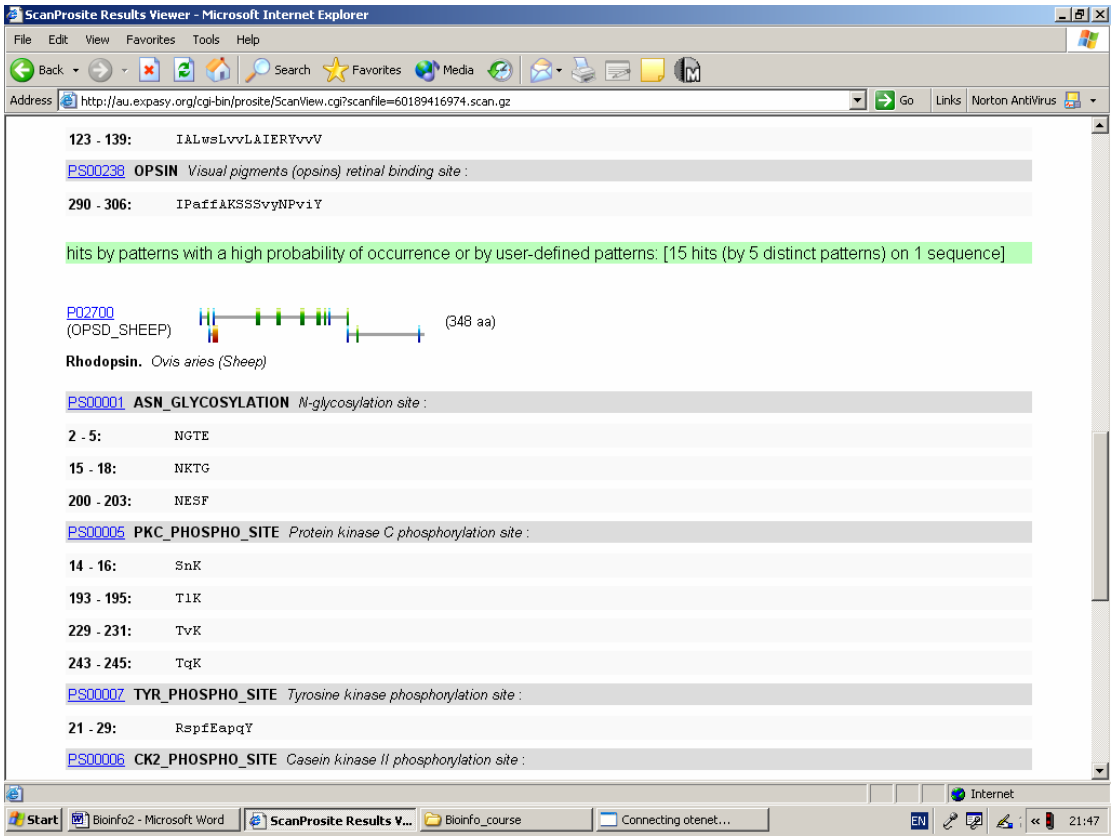

PROSITE: PDOC00211 (documentation file) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/get-prodoc-entry?PDOC00211
Go Links
The attachment site for retinal in the above proteins is a conserved lysine residue in the middle of the seventh transmembrane helix. The pattern we developed includes this residue.
-Consensus pattern: [LIVMFWAC]-[PSGAC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-
[STACP]-x(2)-[DENF]-[AP]-x(2)-[IY]
[K is the retinal binding site]
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in Swiss-Prot: 2.
-Last update: December 2001 / Pattern and text revised.
[1] Applebury M.L., Hargrave P.A.
Vision Res. 26:1881-1895(1986).
[2] Fryxell K.J., Meyerowitz E.M.
J. Mol. Evol. 33:367-378(1991).
[3] Shen D., Jiang M., Hao W., Tao L., Salazar M., Fong H.K.W.
Biochemistry 33:13117-13125(1994).
+----------+
| This PROSITE entry is copyright by the Swiss Institute of Bioinformatics |
| (SIB). There are no restrictions on its use by non-profit institutions as |
| long as its content is in no way modified and this statement is not |
| removed. Usage by and for commercial entities requires a license agreement |
| (See http://www.isb-sib.ch/announce/ or email to license@isb-sib.ch). |
+----------+
(END)
Done
Internet
Start
PROSITE: PDOC002...
Microsoft Word - Document1
5:13 μμ

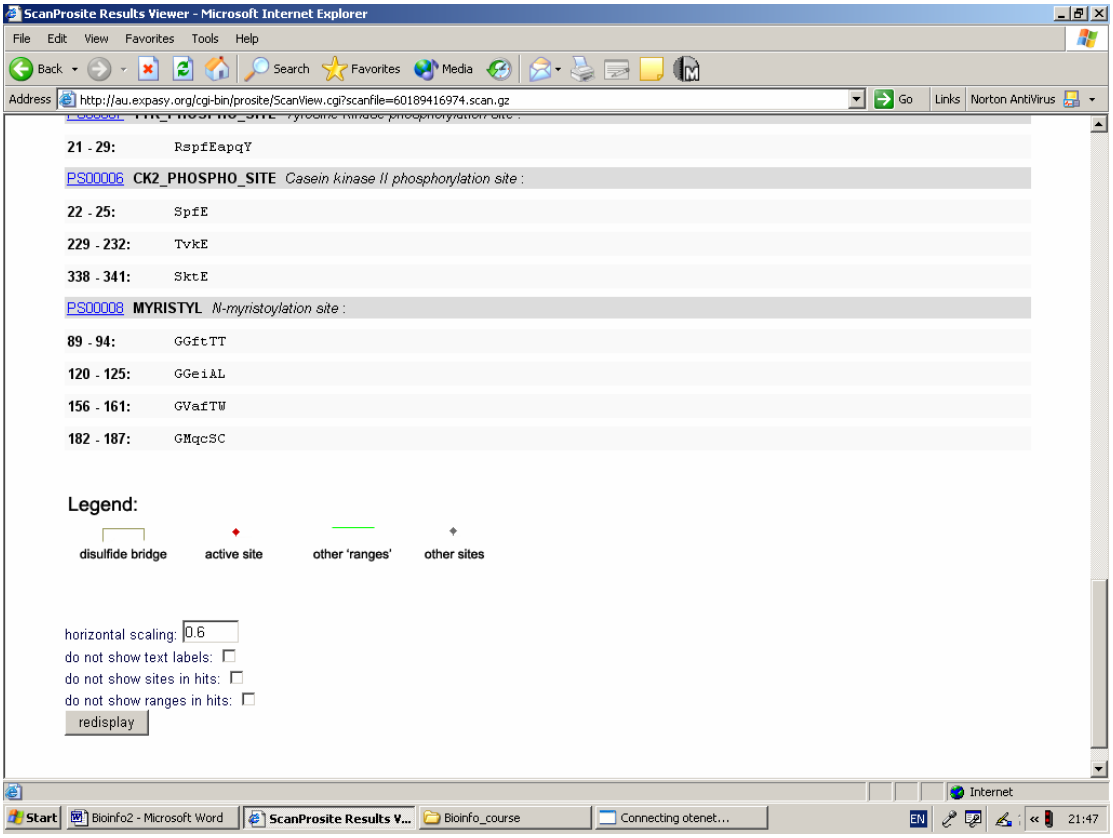
Για τον προσδιορισμό των οικογενειών πρωτεϊνών οι οποίες σχετίζονται με την OPSD_SHEEP, στην πρώτη σελίδα της PROSITE πληκτρολογήστε τον κωδικό (AC) που έχει αυτή στη SWISS-PROT (P02700) στο αντίστοιχο πεδίο και επιλέξτε “Quick Scan”.



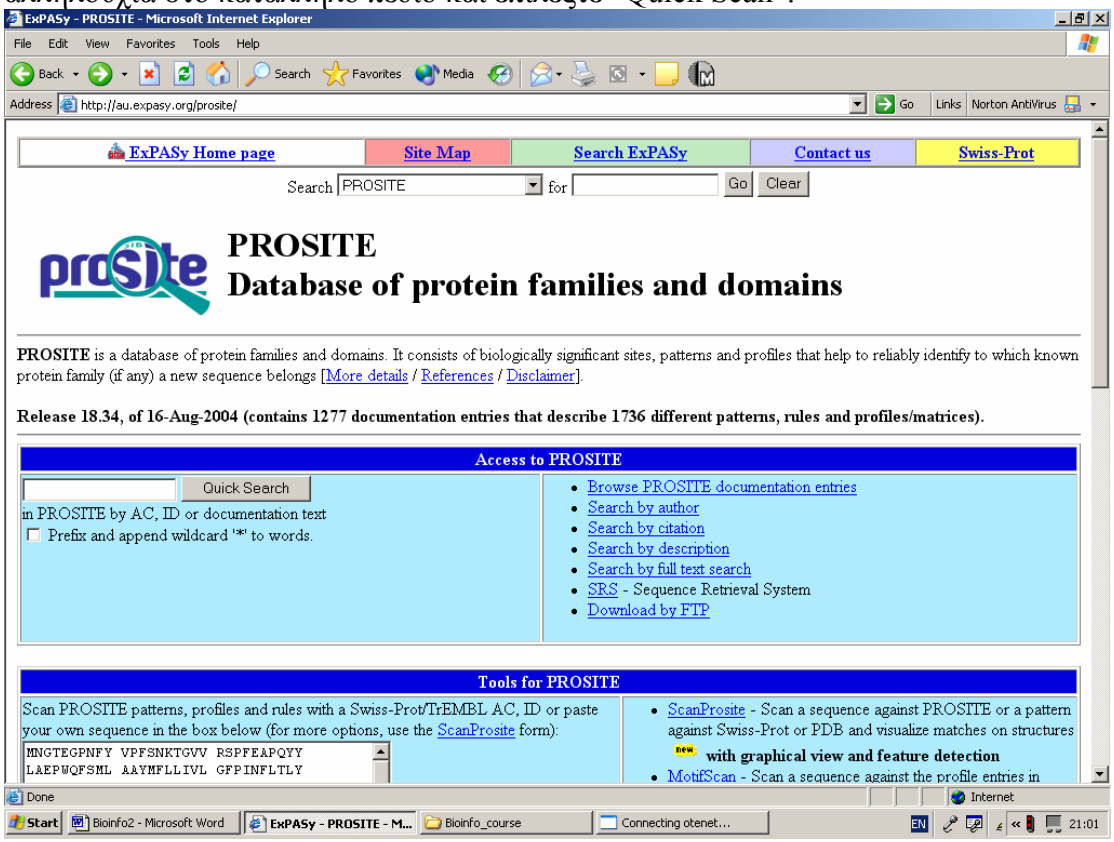


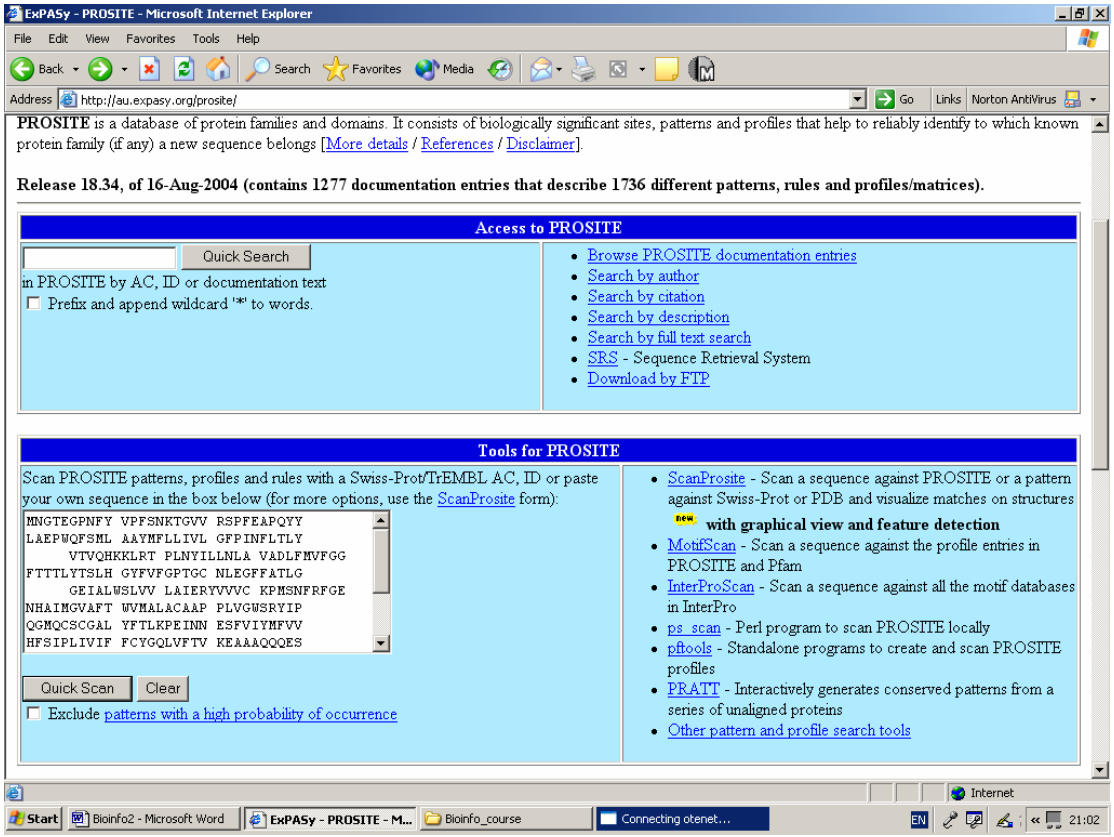

ScanProsite Results Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://au.expasy.org/cgi-bin/prosite/ScanView.cgi?scanfile=60189416974.scan.gz
Go Links Norton AntiVirus
ruler: 1 100 200 300 400 500 600 700 800 900 1000
hits by profiles: [1 hit (by 1 profile) on 1 sequence]
Hits by PS50262 G_PROTEIN_RECEP_F1_2 G-protein coupled receptors family 1 profile :
P02700
(OPSD_SHEEP)
G_PROTEIN_RECEP_F1_2
(348 aa)
Rhodopsin. Ovis aries (Sheep)
54 - 306: score = 41.596
INFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLE
GFFATLGGEIALWSLVVLAIERYVVVCKPMSNFR-FGENHAIMGVAFTWVMALACAAPPL
VG-WSRYIPQGMQCSCGALYFTLkpEINNESFVIYMF-VVHFSIPLIVIFFCYGQLVFTV
KEAAAQQQ--ESATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTH---QGSDFGP
IFMTIPAFFAKSSSVYNPVIY
hits by patterns: [2 hits (by 2 distinct patterns) on 1 sequence]
P02700
(OPSD_SHEEP)
(348 aa)
Rhodopsin. Ovis aries (Sheep)
PS00237 G_PROTEIN_RECEP_F1_1 G-protein coupled receptors family 1 signature :
123 - 139: IALwsLvvLAIERYvvV
PS00238 OPSIN Visual pigments (opsins) retinal binding site :
Internet
Start Bioinfo2 - Microsoft Word ScanProsite Results V... Bioinfo_course Connecting otenet... EN 21:47


ScanProsite Results Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://au.expasy.org/cgi-bin/prosite/ScanView.cgi?scanfile=60189416974.scan.gz
Go Links Norton AntiVirus
123 - 139: IALwsLvvLAIERYvvV
PS00238 OPSIN Visual pigments (opsins) retinal binding site :
290 - 306: IPaffAKSSSvyNPviY
hits by patterns with a high probability of occurrence or by user-defined patterns: [15 hits (by 5 distinct patterns) on 1 sequence]
P02700
(OPSD_SHEEP)
(348 aa)
Rhodopsin. Ovis aries (Sheep)
PS00001 ASN_GLYCOSYLATION N-glycosylation site :
2 - 5: NGTE
15 - 18: NKTG
200 - 203: NESF
PS00005 PKC_PHOSPHO_SITE Protein kinase C phosphorylation site :
14 - 16: SnK
193 - 195: TlK
229 - 231: TvK
243 - 245: TqK
PS00007 TYR_PHOSPHO_SITE Tyrosine kinase phosphorylation site :
21 - 29: RspfEapqY
PS00006 CK2_PHOSPHO_SITE Casein kinase II phosphorylation site :
Internet
Start Bioinfo2 - Microsoft Word ScanProsite Results V... Bioinfo_course Connecting otenet... EN 21:47

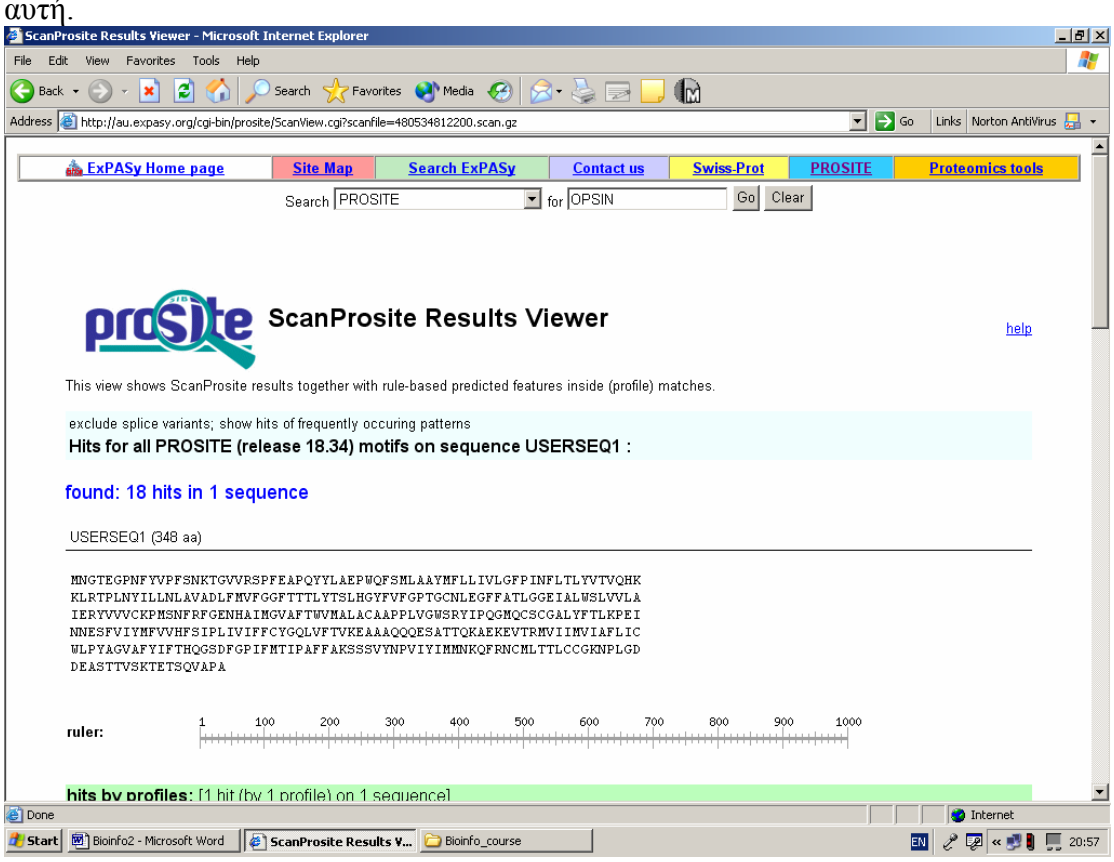

ScanProsite Results Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://au.expasy.org/cgi-bin/prosite/ScanView.cgi?scanfile=60189416974.scan.gz
Go Links Norton AntiVirus
21 - 29: RspfEapqY
PS00006 CK2_PHOSPHO_SITE Casein kinase II phosphorylation site :
22 - 25: SpfE
229 - 232: TvkE
338 - 341: SktE
PS00008 MYRISTYL N-myristoylation site :
89 - 94: GGftTT
120 - 125: GGeiAL
156 - 161: GVafTW
182 - 187: GMqcSC
Legend:
disulfide bridge
active site
other 'ranges'
other sites
horizontal scaling: 0.6
do not show text labels:
do not show sites in hits:
do not show ranges in hits:
redisplay
Internet
Start
Bioinfo2 - Microsoft Word
ScanProsite Results V...
Bioinfo_course
Connecting otenet...
21:47

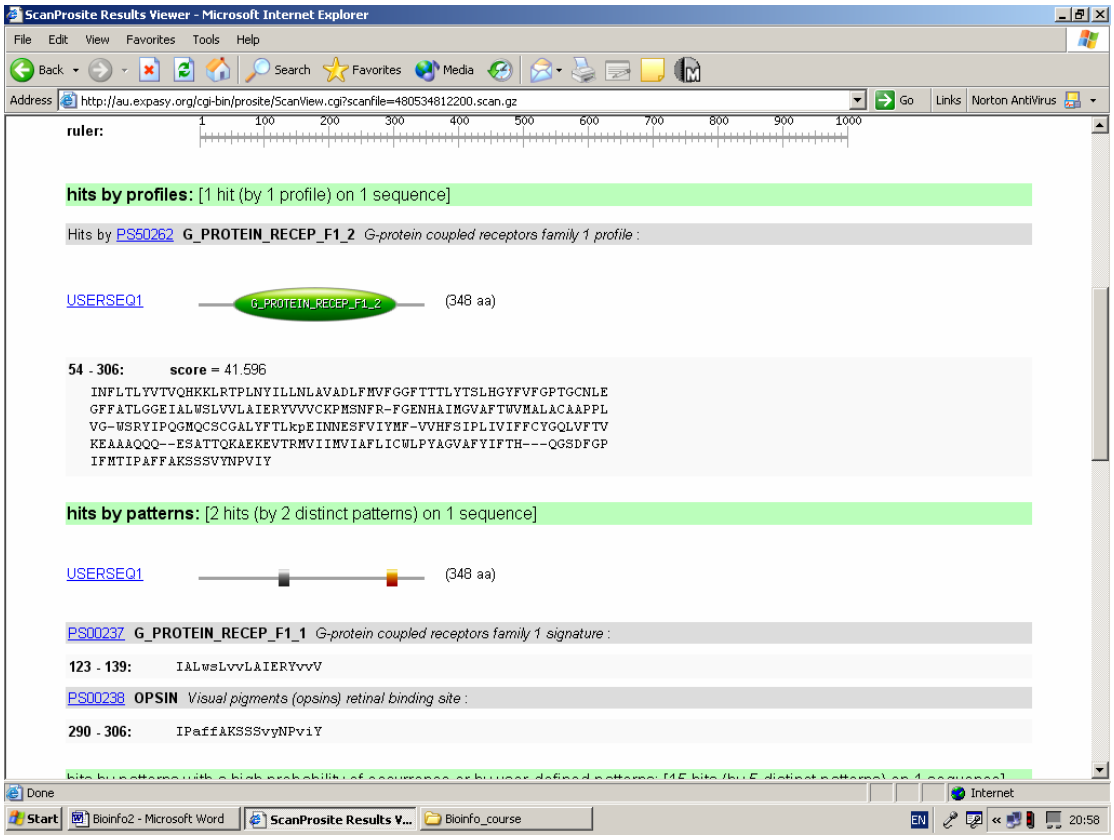
Στην πρώτη σελίδα της PROSITE επίσης, υπάρχει η δυνατότητα για αναζήτηση των πιθανών οικογενειών στις οποίες ενδέχεται να ανήκει μία αλληλουχία, καθώς και της σχετική με την αλληλουχία περιοχής (domain). Για το σκοπό αυτό, εισάγετε την αλληλουχία στο κατάλληλο πεδίο και επιλέξτε “Quick Scan”.

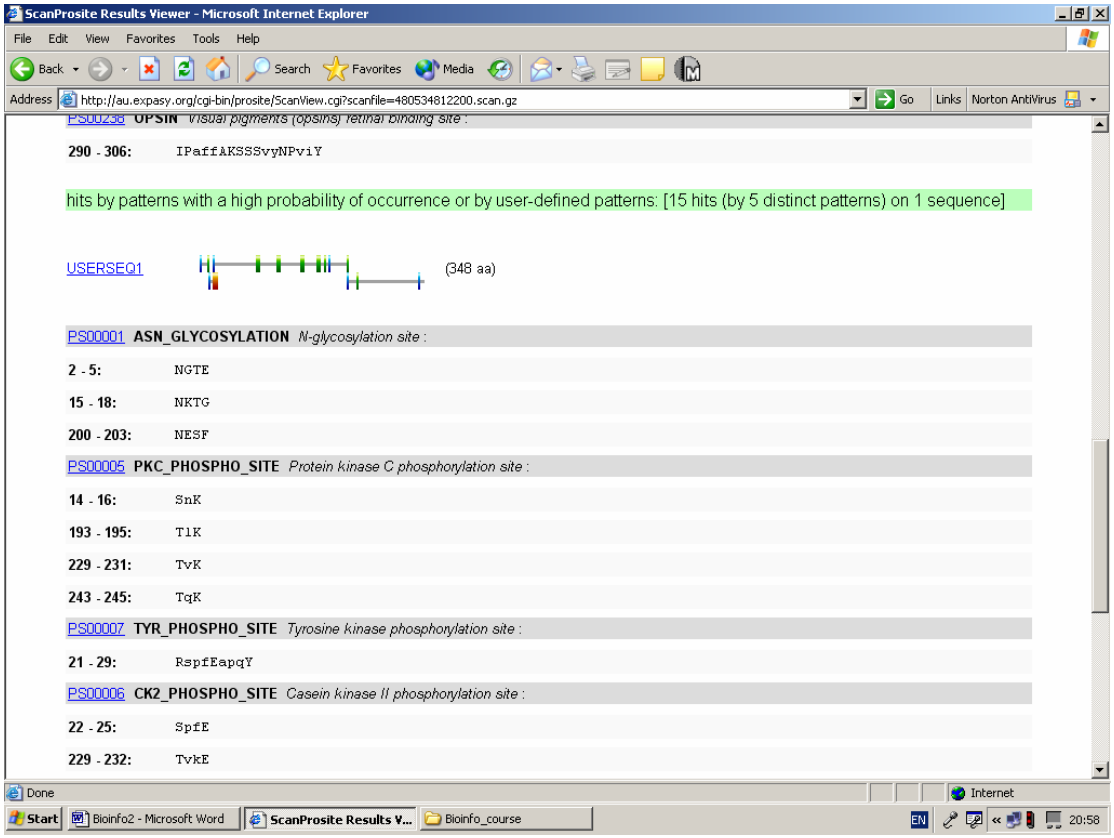


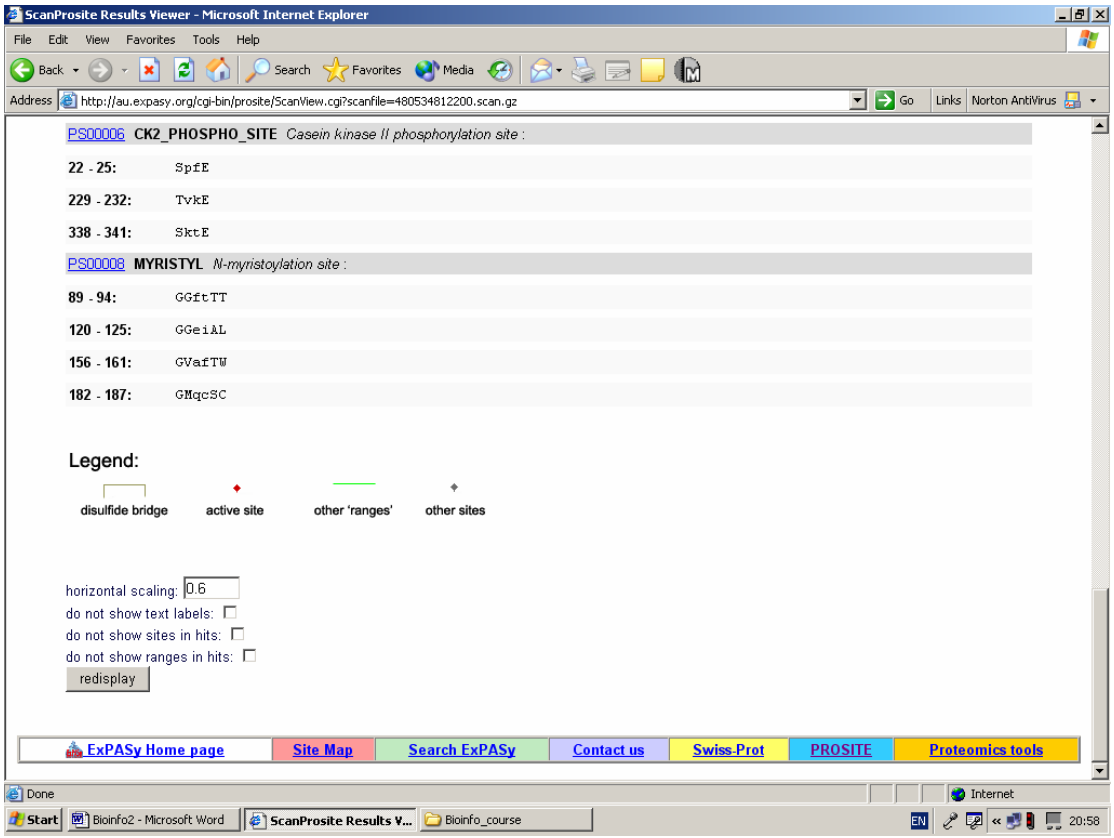


Η PROSITE παρέχει όλα τα δυνατά μοτίβα τα οποία σχετίζονται με την αλληλουχία αυτή.





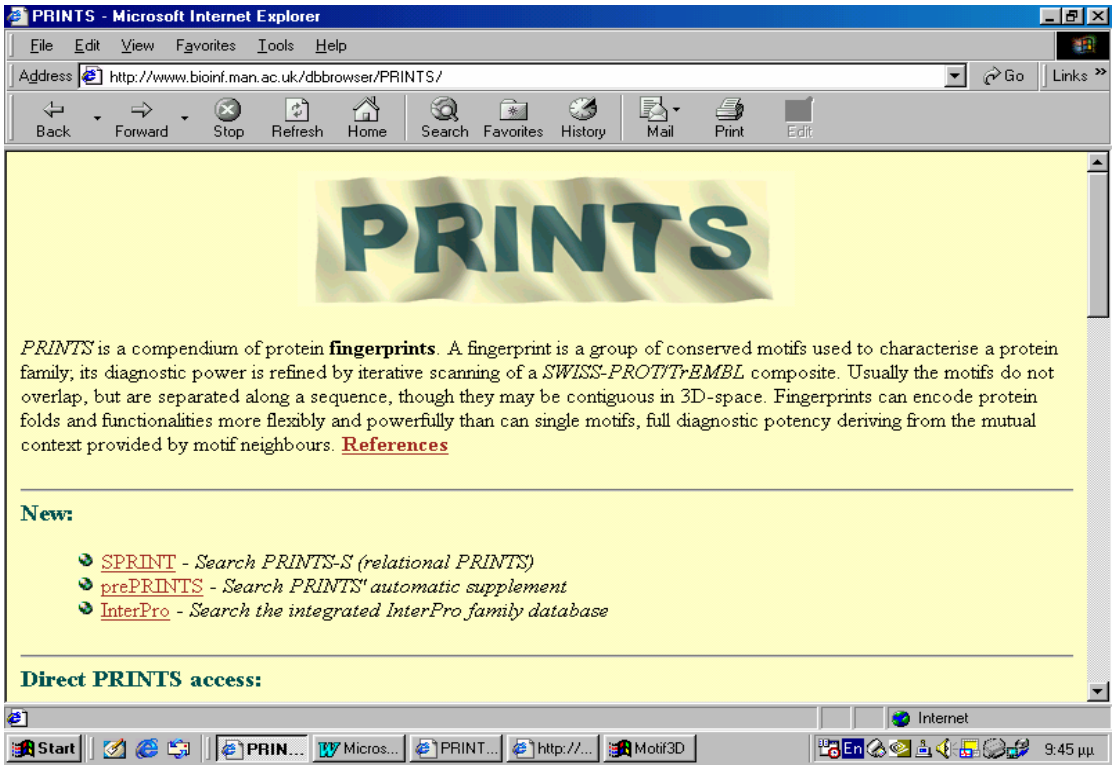

ScanProsite Results Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://au.expasy.org/cgi-bin/prosite/ScanView.cgi?scanfile=480534812200.scan.gz
PS00238 OPSIN Visual pigments (opsins) retinal binding site :
290 - 306: IPaffAKSSSvyNPviY
hits by patterns with a high probability of occurrence or by user-defined patterns: [15 hits (by 5 distinct patterns) on 1 sequence]
USERSEQ1
(348 aa)
PS00001 ASN_GLYCOSYLATION N-glycosylation site :
2 - 5: NGTE
15 - 18: NKTG
200 - 203: NESF
PS00005 PKC_PHOSPHO_SITE Protein kinase C phosphorylation site :
14 - 16: SnK
193 - 195: TlK
229 - 231: TvK
243 - 245: TqK
PS00007 TYR_PHOSPHO_SITE Tyrosine kinase phosphorylation site :
21 - 29: RspfEapqY
PS00006 CK2_PHOSPHO_SITE Casein kinase II phosphorylation site :
22 - 25: SpfE
229 - 232: TvkE
Done
Internet
Start
Bioinfo2 - Microsoft Word
ScanProsite Results V...
Bioinfo_course
20:58


ScanProsite Results Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://au.expasy.org/cgi-bin/prosite/ScanView.cgi?scanfile=480534812200.scan.gz
PS00006 CK2_PHOSPHO_SITE Casein kinase II phosphorylation site :
22 - 25: SpfE
229 - 232: TvkE
338 - 341: SktE
PS00008 MYRISTYL N-myristoylation site :
89 - 94: GGftTT
120 - 125: GGeiAL
156 - 161: GVafTW
182 - 187: GMqcSC
Legend:
disulfide bridge
active site
other 'ranges'
other sites
horizontal scaling: 0.6
do not show text labels:
do not show sites in hits:
do not show ranges in hits:
redisplay
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
PROSITE
Proteomics tools
Done
Internet
Start
Bioinfo2 - Microsoft Word
ScanProsite Results V...
Bioinfo_course
20:58

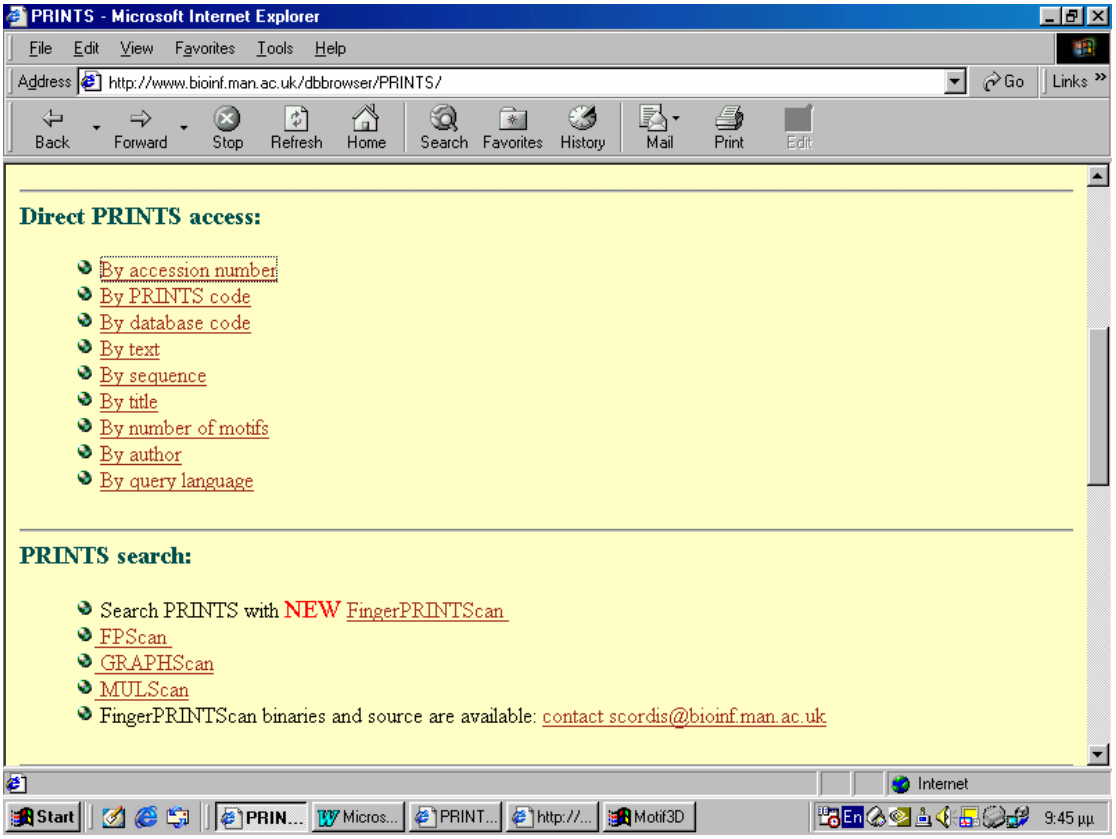
ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

Δευτεροταγείς βάσεις δεδομένων – Αναζήτηση στην PRINTS

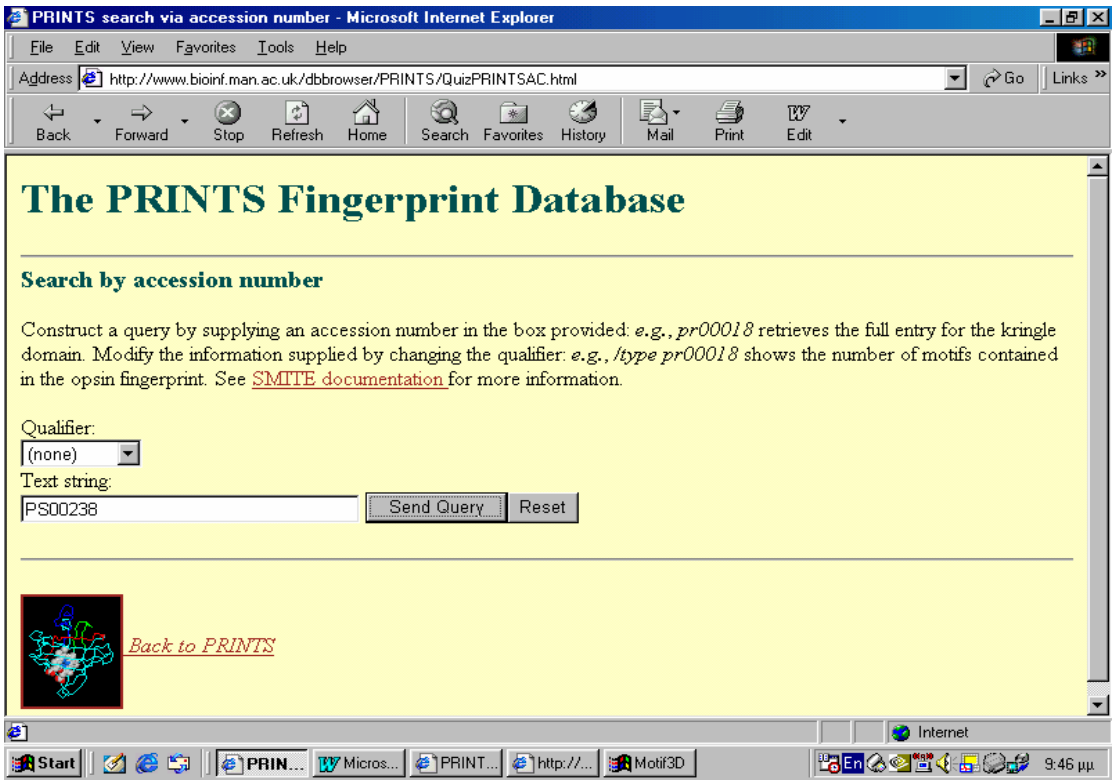
Για πρόσβαση στην PRINTS, στον Internet Explorer πληκτρολογήστε τη διεύθυνση www.bioinf.man.ac/dbbrowser/PRINTS/ και στη συνέχεια επιλέξτε "By accession



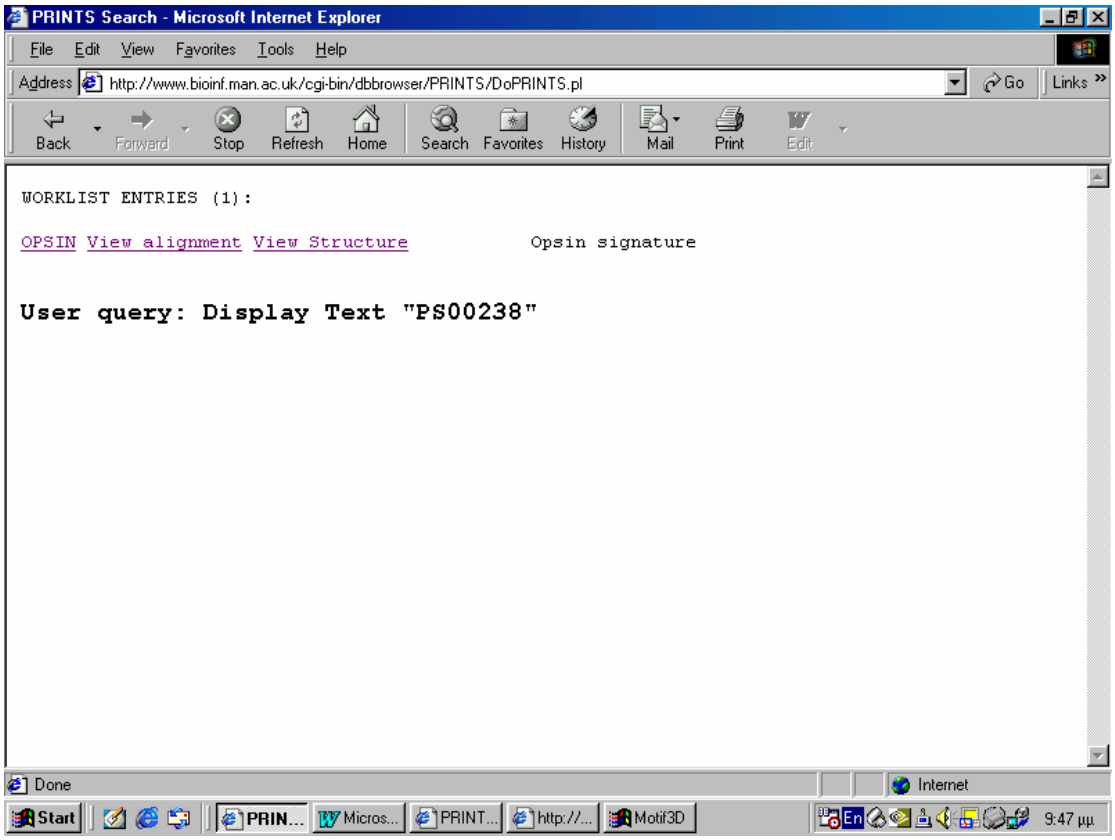
number".



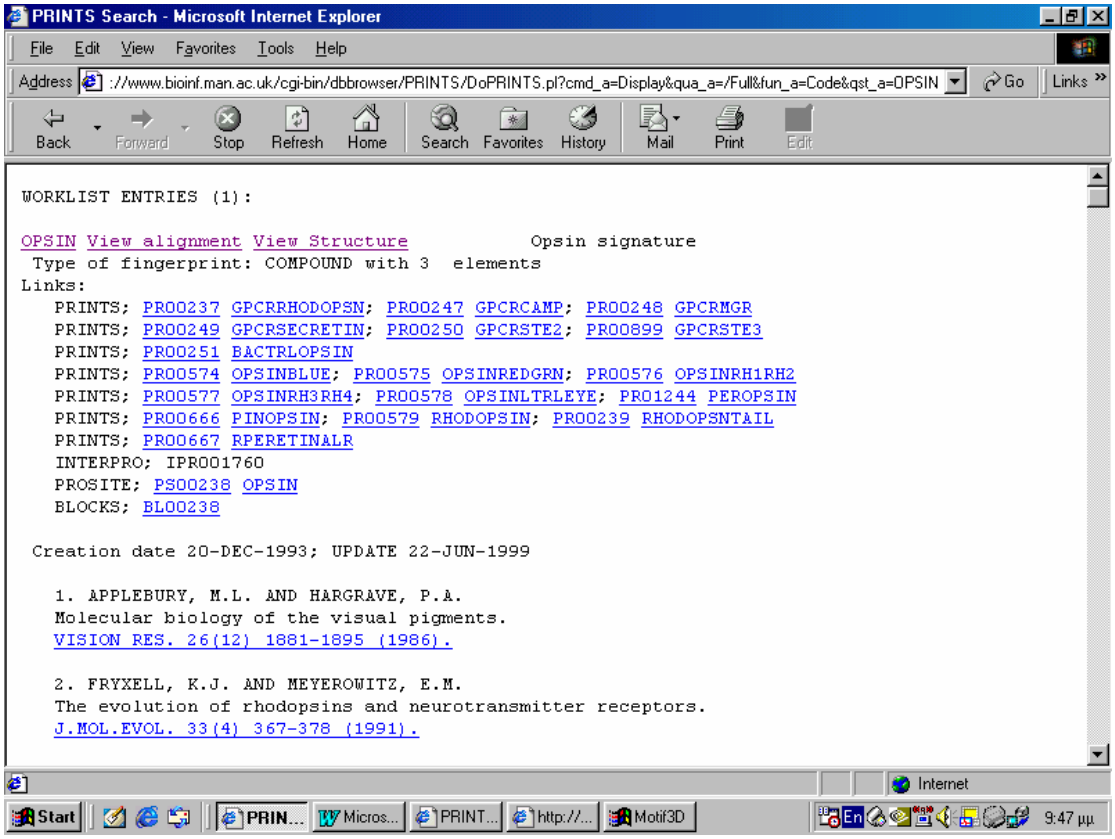
Πληκτρολογήστε σε μορφή κειμένου τον κώδικα της πρωτεΐνης την οποία θέλετε να αναζητήσετε στην PRINTS και επιλέξτε " Send Query ".

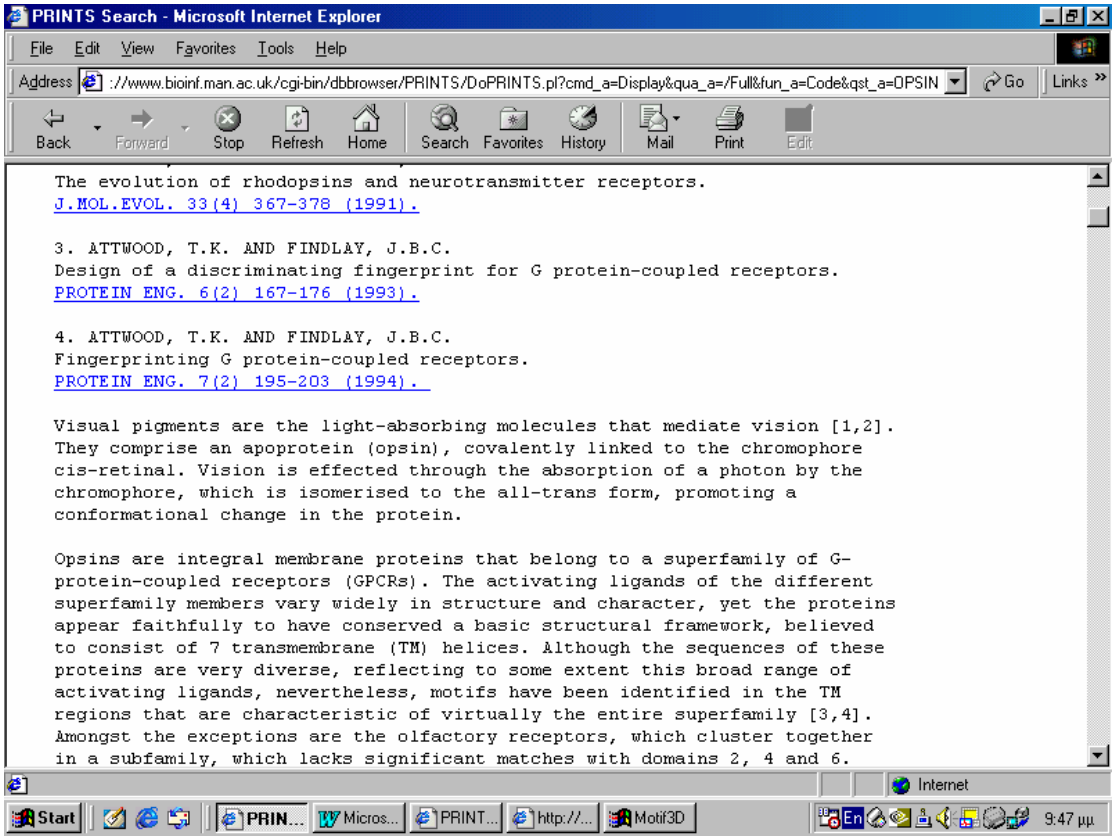


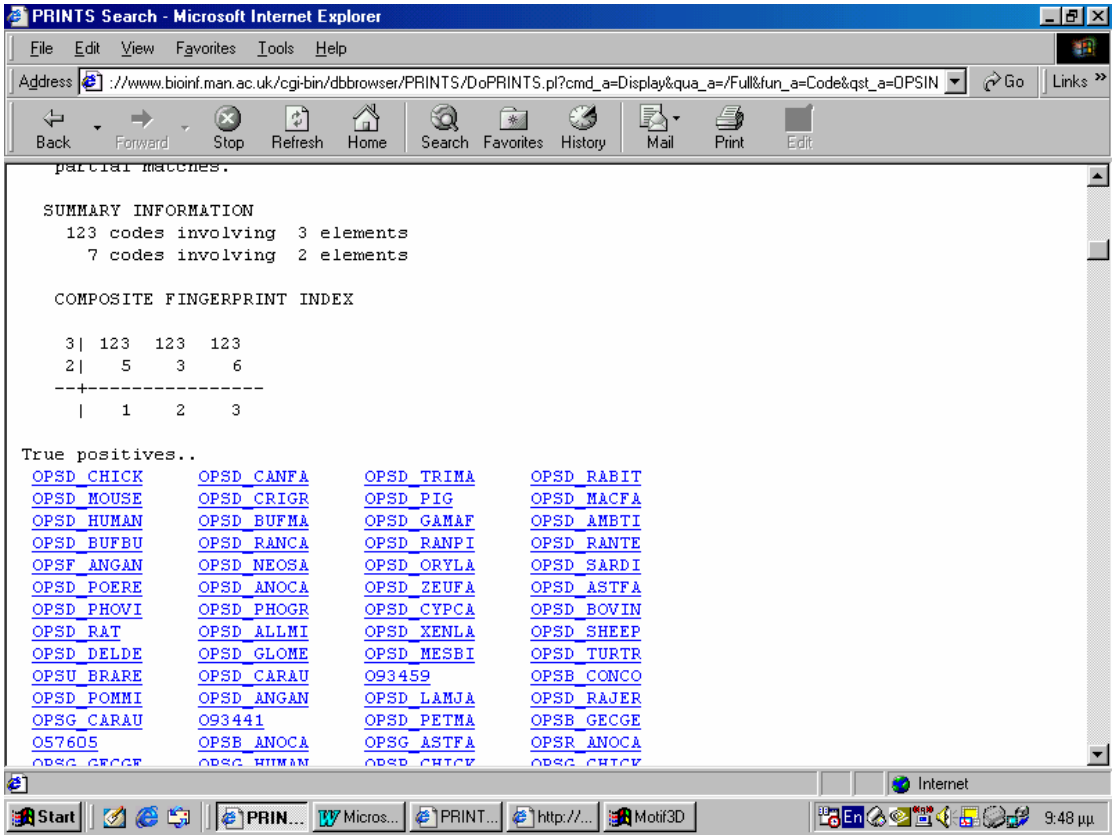
Το PS00238 αντιστοιχεί στην OPSIN. Επιλέξτε την OPSIN για αναζήτηση στην PRINTS με βάση το fingerprint της opsin

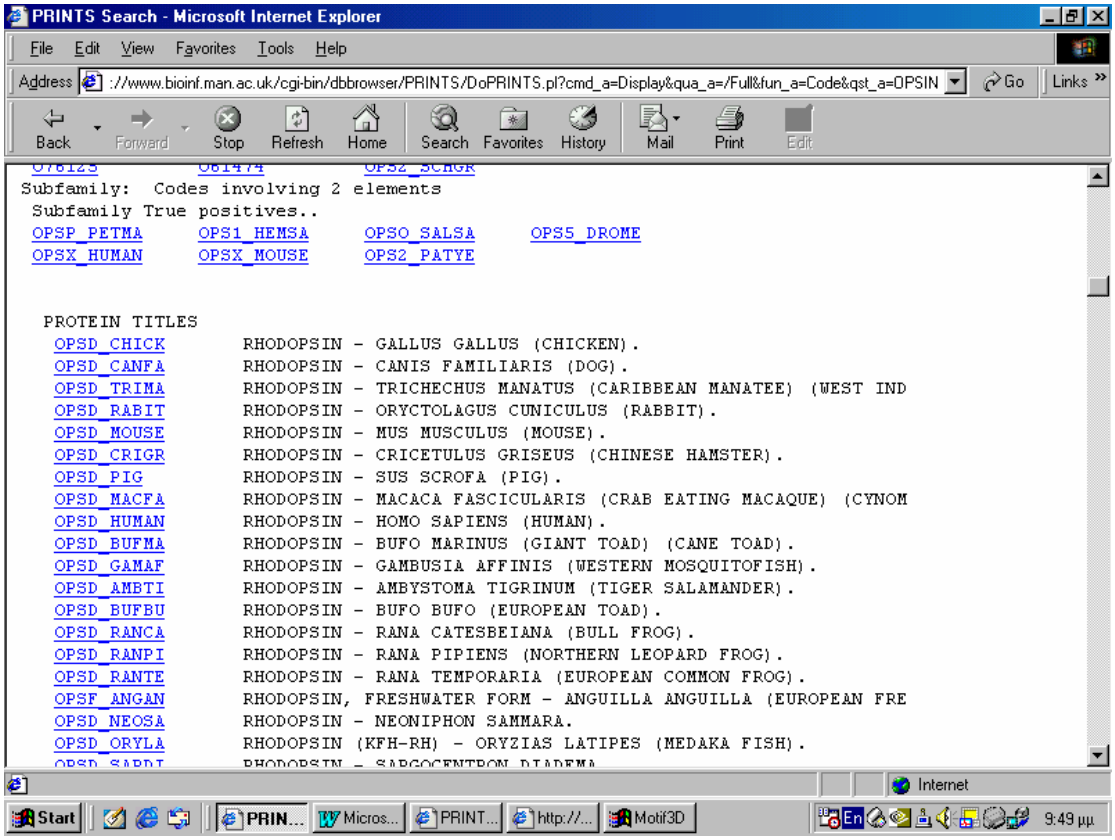


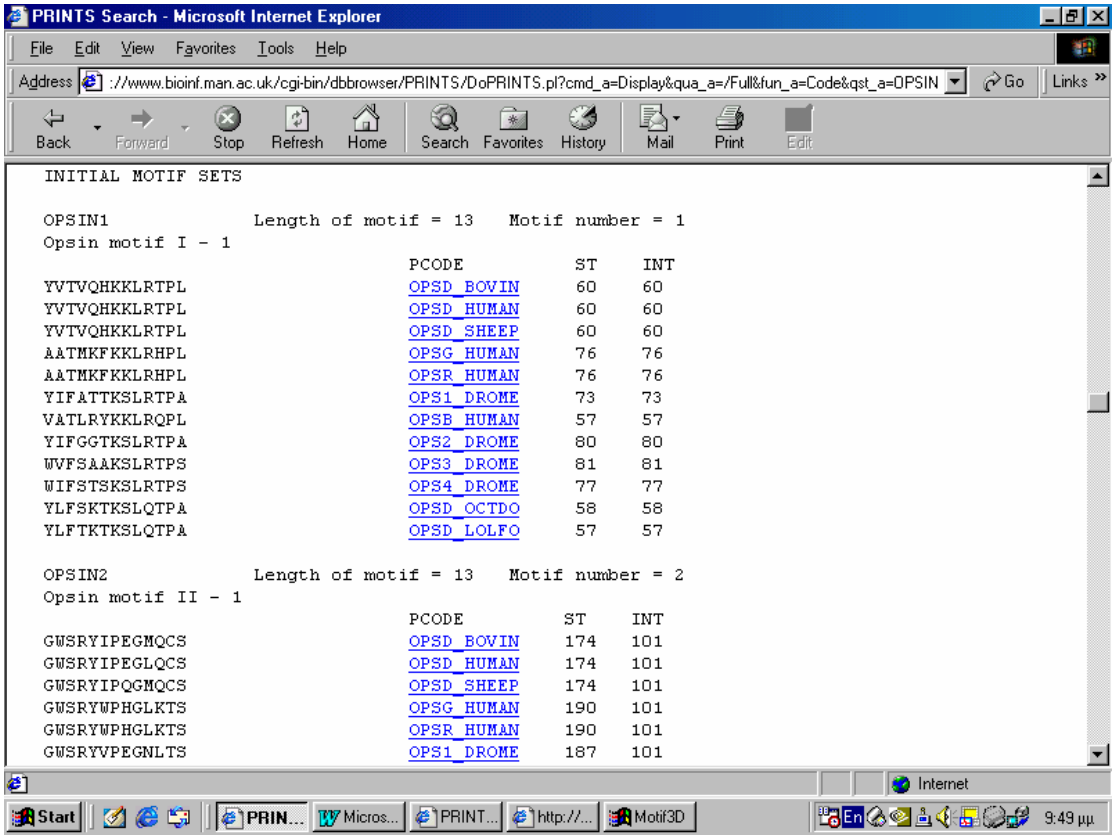
Ως αποτέλεσμα παρουσιάζονται οι συνδέσεις (cross-links), η περιγραφή της opsin, η διαγνωστική επίδοση / αξία και τα μοτίβα.

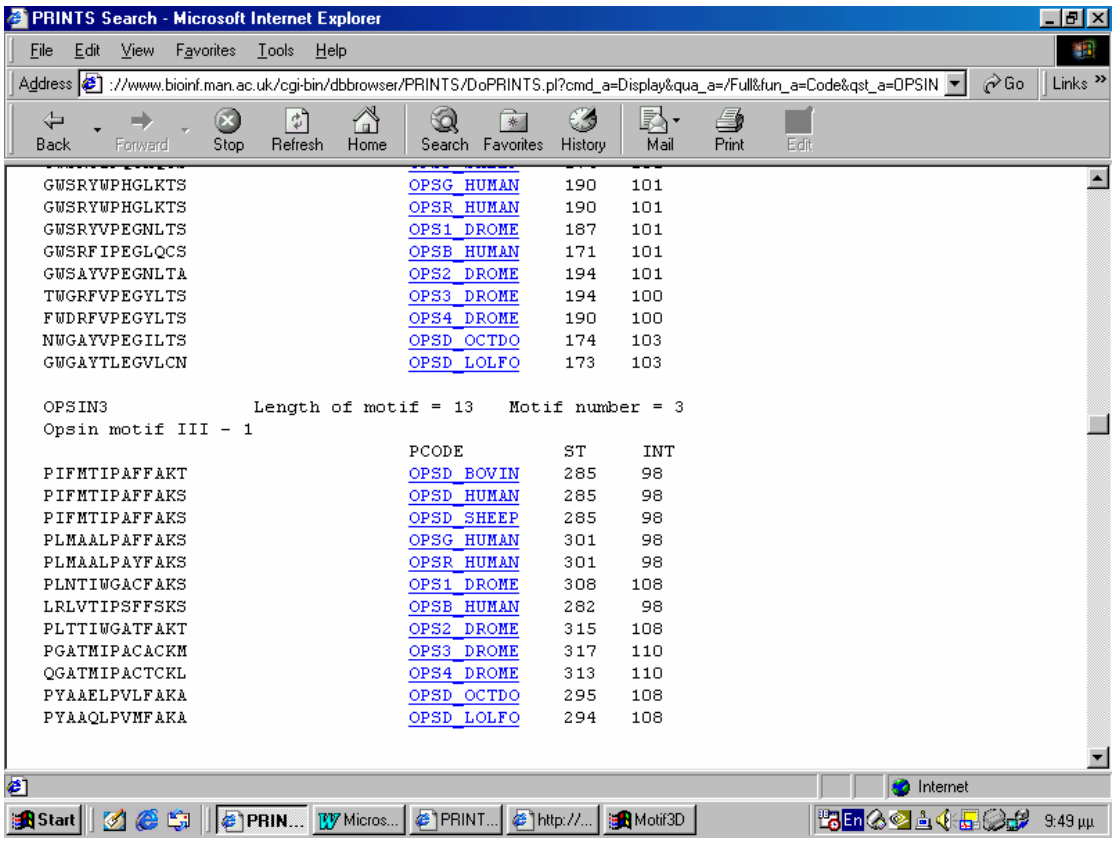

PRINTS Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address ://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/PRINTS/DoPRINTS.pl?cmd_a=Display&qua_a=/Full&fun_a=Code&qst_a=OPSIN
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
WORKLIST ENTRIES (1):
OPSIN View alignment View Structure Opsin signature
Type of fingerprint: COMPOUND with 3 elements
Links:
PRINTS; PR00237 GPCRRHODOPSN; PR00247 GPCRCAMP; PR00248 GPCRMGR
PRINTS; PR00249 GPCRSECRETIN; PR00250 GPCRSTE2; PR00899 GPCRSTE3
PRINTS; PR00251 BACTRLOPSIN
PRINTS; PR00574 OPSINBLUE; PR00575 OPSINREDGRN; PR00576 OPSINRH1RH2
PRINTS; PR00577 OPSINRH3RH4; PR00578 OPSINLTRLEYE; PR01244 PEROPSIN
PRINTS; PR00666 PINOPSIN; PR00579 RHODOPSIN; PR00239 RHODOPSNTAIL
PRINTS; PR00667 RPERETINALR
INTERPRO; IPR001760
PROSITE; PS00238 OPSIN
BLOCKS; BL00238
Creation date 20-DEC-1993; UPDATE 22-JUN-1999
1. APPLEBURY, M.L. AND HARGRAVE, P.A.
Molecular biology of the visual pigments.
VISION RES. 26(12) 1881-1895 (1986).
2. FRYXELL, K.J. AND MEYEROWITZ, E.M.
The evolution of rhodopsins and neurotransmitter receptors.
J.MOL.EVOL. 33(4) 367-378 (1991).
Internet
Start PRIN... Micros... PRINT... http://... Motif3D 9:47 µµ

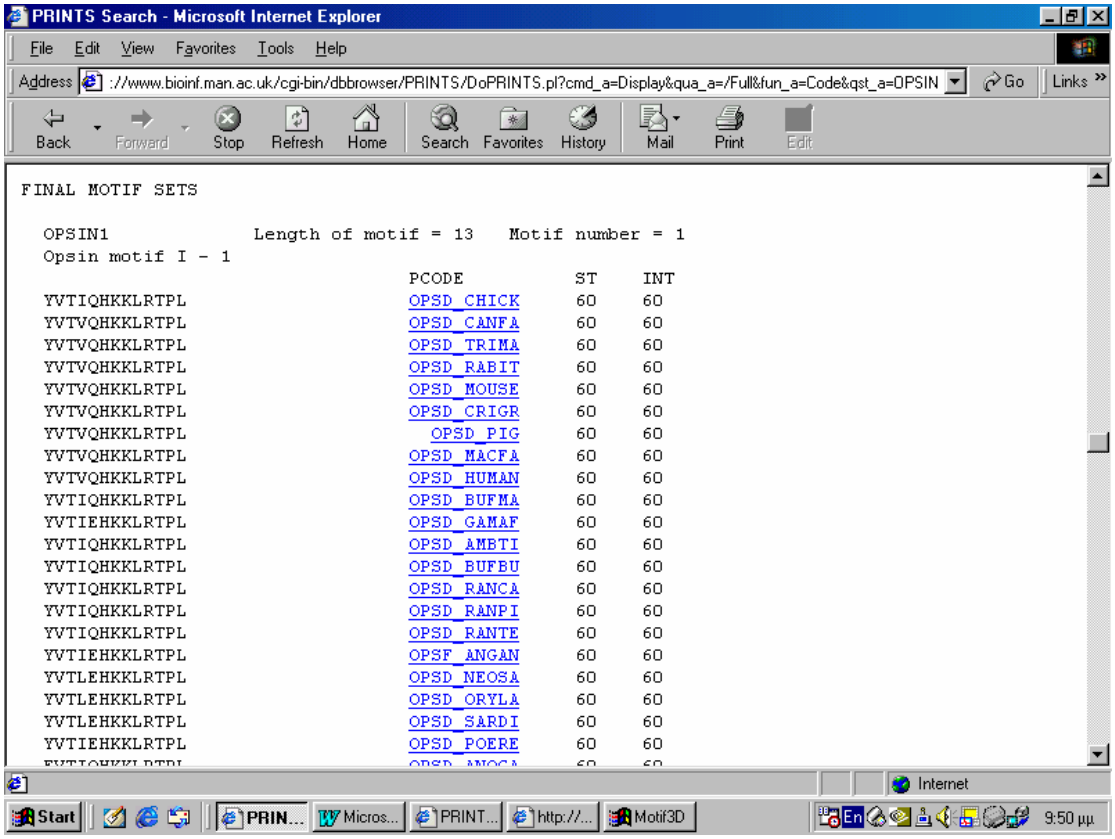

PRINTS Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address ://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/PRINTS/DoPRINTS.pl?cmd_a=Display&qua_a=/Full&fun_a=Code&qst_a=OPSIN
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
The evolution of rhodopsins and neurotransmitter receptors.
J.MOL.EVOL. 33(4) 367-378 (1991).
3. ATTWOOD, T.K. AND FINDLAY, J.B.C.
Design of a discriminating fingerprint for G protein-coupled receptors.
PROTEIN ENG. 6(2) 167-176 (1993).
4. ATTWOOD, T.K. AND FINDLAY, J.B.C.
Fingerprinting G protein-coupled receptors.
PROTEIN ENG. 7(2) 195-203 (1994).
Visual pigments are the light-absorbing molecules that mediate vision [1,2].
They comprise an apoprotein (opsin), covalently linked to the chromophore
cis-retinal. Vision is effected through the absorption of a photon by the
chromophore, which is isomerised to the all-trans form, promoting a
conformational change in the protein.
Opsins are integral membrane proteins that belong to a superfamily of G-
protein-coupled receptors (GPCRs). The activating ligands of the different
superfamily members vary widely in structure and character, yet the proteins
appear faithfully to have conserved a basic structural framework, believed
to consist of 7 transmembrane (TM) helices. Although the sequences of these
proteins are very diverse, reflecting to some extent this broad range of
activating ligands, nevertheless, motifs have been identified in the TM
regions that are characteristic of virtually the entire superfamily [3,4].
Amongst the exceptions are the olfactory receptors, which cluster together
in a subfamily, which lacks significant matches with domains 2, 4 and 6.
Internet
Start PRIN... Micros... PRINT... http://... Motif3D 9:47 µµ

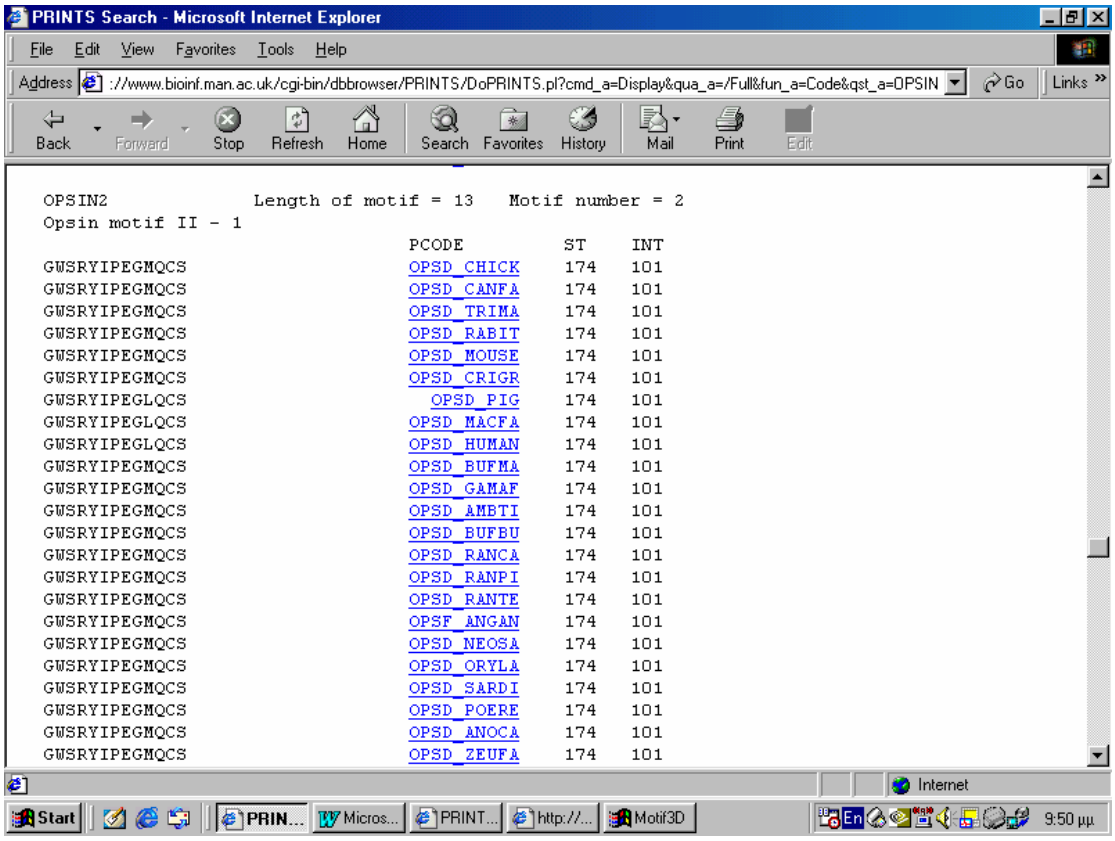

PRINTS Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address ://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/PRINTS/DoPRINTS.pl?cmd_a=Display&qua_a=/Full&fun_a=Code&qst_a=OPSIN
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
partial matches.
SUMMARY INFORMATION
123 codes involving 3 elements
7 codes involving 2 elements
COMPOSITE FINGERPRINT INDEX
3| 123 123 123
2| 5 3 6
--+---
| 1 2 3
True positives..
OPSD_CHICK OPSD_CANFA OPSD_TRIMA OPSD_RABIT
OPSD_MOUSE OPSD_CRIGR OPSD_PIG OPSD_MACFA
OPSD_HUMAN OPSD_BUFMA OPSD_GAMAF OPSD_AMBTI
OPSD_BUFBU OPSD_RANCA OPSD_RANPI OPSD_RANTE
OPSF_ANGAN OPSD_NEOSA OPSD_ORYLA OPSD_SARDI
OPSD_POERE OPSD_ANOCA OPSD_ZEUFA OPSD_ASTFA
OPSD_PHOVI OPSD_PHOGR OPSD_CYPCA OPSD_BOVIN
OPSD_RAT OPSD_ALLMI OPSD_XENLA OPSD_SHEEP
OPSD_DELDE OPSD_GLOME OPSD_MESBI OPSD_TURTR
OPSU_BRARE OPSD_CARAU O93459 OPSB_CONCO
OPSD_POMMI OPSD_ANGAN OPSD_LAMJA OPSD_RAJER
OPSG_CARAU O93441 OPSD_PETMA OPSB_GECGE
O57605 OPSB_ANOCA OPSG_ASTFA OPSR_ANOCA
Internet
Start PRIN... Micros... PRINT... http://... Motif3D 9:48 µµ

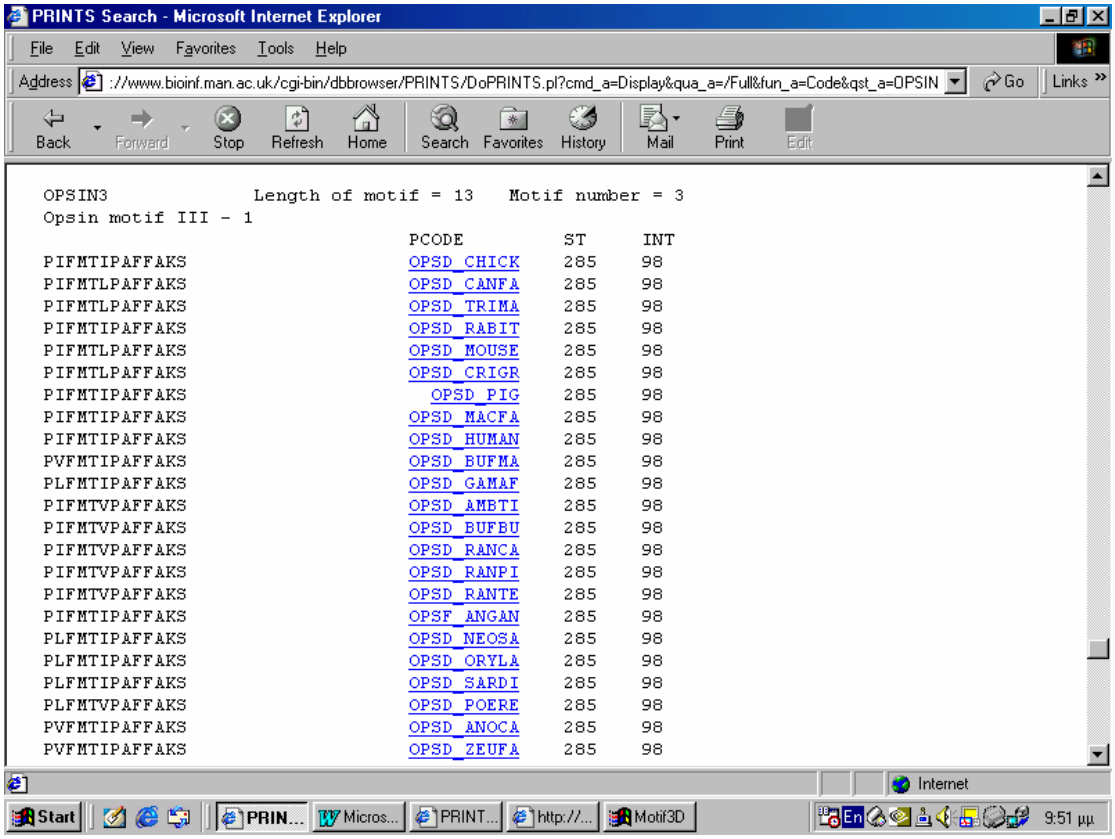

PRINTS Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address ://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/PRINTS/DoPRINTS.pl?cmd_a=Display&qua_a=/Full&fun_a=Code&qst_a=OPSIN
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Subfamily: Codes involving 2 elements
Subfamily True positives..
OPSP_PETMA OPS1_HEMSA OPSO_SALSA OPS5_DROME
OPSX_HUMAN OPSX_MOUSE OPS2_PATYE
PROTEIN TITLES
OPSD_CHICK RHODOPSIN - GALLUS GALLUS (CHICKEN).
OPSD_CANFA RHODOPSIN - CANIS FAMILIARIS (DOG).
OPSD_TRIMA RHODOPSIN - TRICHECHUS MANATUS (CARIBBEAN MANATEE) (WEST IND
OPSD_RABIT RHODOPSIN - ORYCTOLAGUS CUNICULUS (RABBIT).
OPSD_MOUSE RHODOPSIN - MUS MUSCULUS (MOUSE).
OPSD_CRIGR RHODOPSIN - CRICETULUS GRISEUS (CHINESE HAMSTER).
OPSD_PIG RHODOPSIN - SUS SCROFA (PIG).
OPSD_MACFA RHODOPSIN - MACACA FASCICULARIS (CRAB EATING MACAQUE) (CYNOM
OPSD_HUMAN RHODOPSIN - HOMO SAPIENS (HUMAN).
OPSD_BUFMA RHODOPSIN - BUFO MARINUS (GIANT TOAD) (CANE TOAD).
OPSD_GAMAF RHODOPSIN - GAMBUSIA AFFINIS (WESTERN MOSQUITOFISH).
OPSD_AMBTI RHODOPSIN - AMBYSTOMA TIGRINUM (TIGER SALAMANDER).
OPSD_BUFBU RHODOPSIN - BUFO BUFO (EUROPEAN TOAD).
OPSD_RANCA RHODOPSIN - RANA CATESBEIANA (BULL FROG).
OPSD_RANPI RHODOPSIN - RANA PIPIENS (NORTHERN LEOPARD FROG).
OPSD_RANTE RHODOPSIN - RANA TEMPORARIA (EUROPEAN COMMON FROG).
OPSF_ANGAN RHODOPSIN, FRESHWATER FORM - ANGUILLA ANGUILLA (EUROPEAN FRE
OPSD_NEOSA RHODOPSIN - NEONIPHON SAMMARA.
OPSD_ORYLA RHODOPSIN (KFH-RH) - ORYZIAS LATIPES (MEDAKA FISH).
Internet
Start PRIN... Micros... PRINT... http://... Motif3D 9:49 µµ


PRINTS Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address ://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/PRINTS/DoPRINTS.pl?cmd_a=Display&qua_a=/Full&fun_a=Code&qst_a=OPSIN
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
INITIAL MOTIF SETS
OPSIN1 Length of motif = 13 Motif number = 1
Opsin motif I - 1
PCODE ST INT
YVTVQHKKLRTPL OPSD_BOVIN 60 60
YVTVQHKKLRTPL OPSD_HUMAN 60 60
YVTVQHKKLRTPL OPSD_SHEEP 60 60
AATMKFKKLRHPL OPSG_HUMAN 76 76
AATMKFKKLRHPL OPSR_HUMAN 76 76
YIFATTKSLRTPA OPS1_DROME 73 73
VATLRYKKLRQPL OPSB_HUMAN 57 57
YIFGGTKSLRTPA OPS2_DROME 80 80
WVFSAAKSLRTPS OPS3_DROME 81 81
WIFSTSKSLRTPS OPS4_DROME 77 77
YLFSKTKSLQTPA OPSD_OCTDO 58 58
YLFTKTKSLQTPA OPSD_LOLFO 57 57
OPSIN2 Length of motif = 13 Motif number = 2
Opsin motif II - 1
PCODE ST INT
GWSRYIPEGMQCS OPSD_BOVIN 174 101
GWSRYIPEGLQCS OPSD_HUMAN 174 101
GWSRYIPQGMQCS OPSD_SHEEP 174 101
GWSRYWPHGLKTS OPSG_HUMAN 190 101
GWSRYWPHGLKTS OPSR_HUMAN 190 101
GWSRYVPEGNLTS OPS1_DROME 187 101
Internet
Start PRIN... Micros... PRINT... http://... Motif3D 9:49 µµ

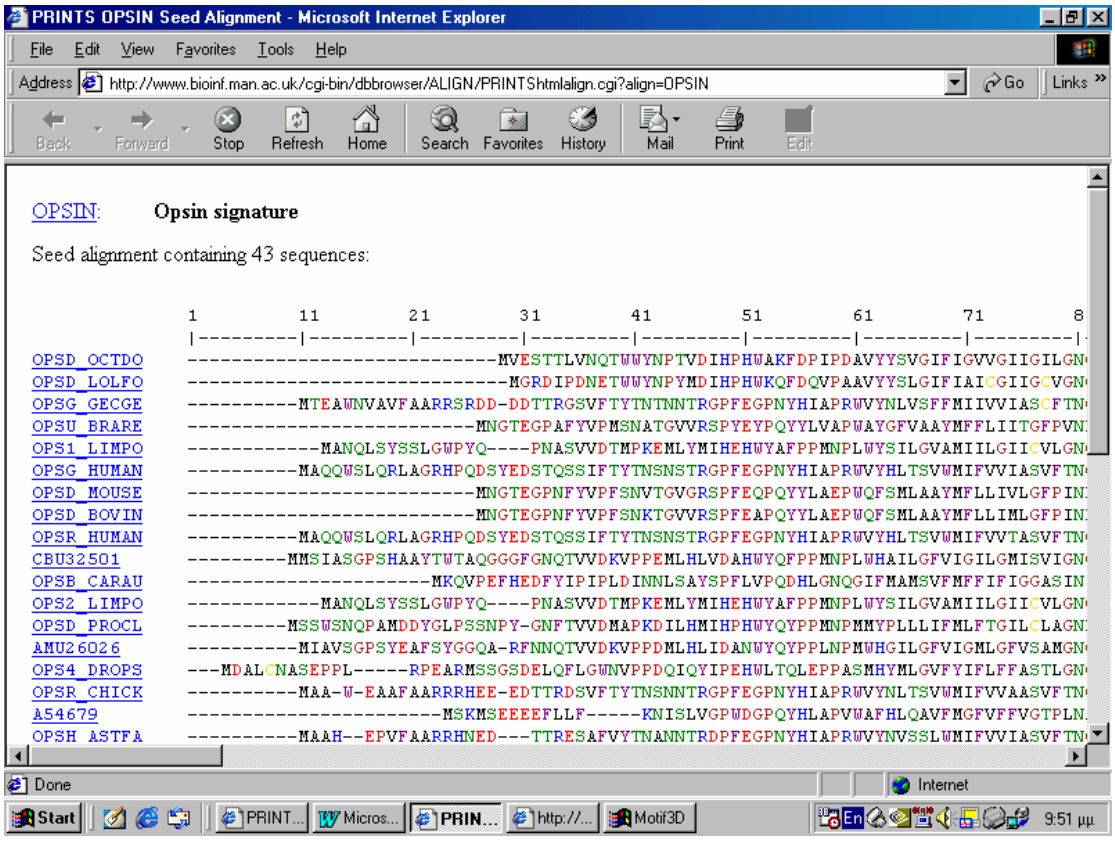

PRINTS Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address ://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/PRINTS/DoPRINTS.pl?cmd_a=Display&qua_a=/Full&fun_a=Code&qst_a=OPSIN
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
GWSRYWPHGLKTS OPSG_HUMAN 190 101
GWSRYWPHGLKTS OPSR_HUMAN 190 101
GWSRYVPEGNLTS OPS1_DROME 187 101
GWSRFIPEGLQCS OPSB_HUMAN 171 101
GWSAYVPEGNLTA OPS2_DROME 194 101
TWGRFVPEGYLTS OPS3_DROME 194 100
FWDRFVPEGYLTS OPS4_DROME 190 100
NWGAYVPEGILTS OPSD_OCTDO 174 103
GWGAYTLEGVLCN OPSD_LOLFO 173 103
OPSIN3 Length of motif = 13 Motif number = 3
Opsin motif III - 1
PCODE ST INT
PIFMTIPAFFAKT OPSD_BOVIN 285 98
PIFMTIPAFFAKS OPSD_HUMAN 285 98
PIFMTIPAFFAKS OPSD_SHEEP 285 98
PLMAALPAFFAKS OPSG_HUMAN 301 98
PLMAALPAYFAKS OPSR_HUMAN 301 98
PLNTIWGACFAKS OPS1_DROME 308 108
LRLVTIPSFFSKS OPSB_HUMAN 282 98
PLTTIWGATFAKT OPS2_DROME 315 108
PGATMIPACACKM OPS3_DROME 317 110
QGATMIPACTCKL OPS4_DROME 313 110
PYAAELPVLFAKA OPSD_OCTDO 295 108
PYAAQLPVMFAKA OPSD_LOLFO 294 108
Internet
Start PRIN... Micros... PRINT... http://... Motif3D 9:49 µµ


PRINTS Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address ://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/PRINTS/DoPRINTS.pl?cmd_a=Display&qua_a=/Full&fun_a=Code&qst_a=OPSIN
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
FINAL MOTIF SETS
OPSIN1 Length of motif = 13 Motif number = 1
Opsin motif I - 1
PCODE ST INT
YVTIQHKKLRTPL OPSD_CHICK 60 60
YVTVQHKKLRTPL OPSD_CANFA 60 60
YVTVQHKKLRTPL OPSD_TRIMA 60 60
YVTVQHKKLRTPL OPSD_RABIT 60 60
YVTVQHKKLRTPL OPSD_MOUSE 60 60
YVTVQHKKLRTPL OPSD_CRIGR 60 60
YVTVQHKKLRTPL OPSD_PIG 60 60
YVTVQHKKLRTPL OPSD_MACFA 60 60
YVTVQHKKLRTPL OPSD_HUMAN 60 60
YVTIQHKKLRTPL OPSD_BUFMA 60 60
YVTIEHKKLRTPL OPSD_GAMAF 60 60
YVTIQHKKLRTPL OPSD_AMBTI 60 60
YVTIQHKKLRTPL OPSD_BUFBU 60 60
YVTIQHKKLRTPL OPSD_RANCA 60 60
YVTIQHKKLRTPL OPSD_RANPI 60 60
YVTIQHKKLRTPL OPSD_RANTE 60 60
YVTIEHKKLRTPL OPSF_ANGAN 60 60
YVTLEHKKLRTPL OPSD_NEOSA 60 60
YVTLEHKKLRTPL OPSD_ORYLA 60 60
YVTLEHKKLRTPL OPSD_SARDI 60 60
YVTIEHKKLRTPL OPSD_POERE 60 60
Internet
Start PRIN... Micros... PRINT... http://... Motif3D 9:50 µµ

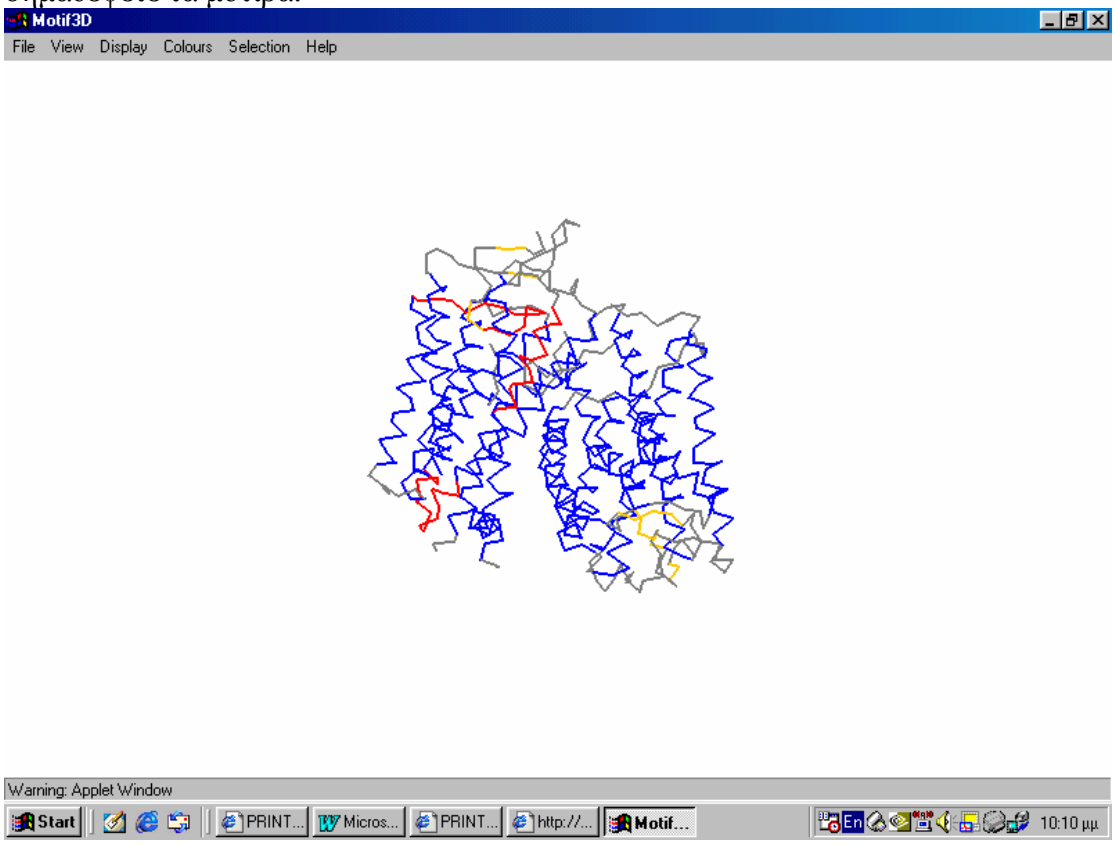

PRINTS Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address ://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/PRINTS/DoPRINTS.pl?cmd_a=Display&qua_a=/Full&fun_a=Code&qst_a=OPSIN
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
OPSIN2 Length of motif = 13 Motif number = 2
Opsin motif II - 1
PCODE ST INT
GWSRYIPEGMQCS OPSD_CHICK 174 101
GWSRYIPEGMQCS OPSD_CANFA 174 101
GWSRYIPEGMQCS OPSD_TRIMA 174 101
GWSRYIPEGMQCS OPSD_RABIT 174 101
GWSRYIPEGMQCS OPSD_MOUSE 174 101
GWSRYIPEGMQCS OPSD_CRIGR 174 101
GWSRYIPEGLQCS OPSD_PIG 174 101
GWSRYIPEGLQCS OPSD_MACFA 174 101
GWSRYIPEGLQCS OPSD_HUMAN 174 101
GWSRYIPEGMQCS OPSD_BUFMA 174 101
GWSRYIPEGMQCS OPSD_GAMAF 174 101
GWSRYIPEGMQCS OPSD_AMBTI 174 101
GWSRYIPEGMQCS OPSD_BUFBU 174 101
GWSRYIPEGMQCS OPSD_RANCA 174 101
GWSRYIPEGMQCS OPSD_RANPI 174 101
GWSRYIPEGMQCS OPSD_RANTE 174 101
GWSRYIPEGMQCS OPSF_ANGAN 174 101
GWSRYIPEGMQCS OPSD_NEOSA 174 101
GWSRYIPEGMQCS OPSD_ORYLA 174 101
GWSRYIPEGMQCS OPSD_SARDI 174 101
GWSRYIPEGMQCS OPSD_POERE 174 101
GWSRYIPEGMQCS OPSD_ANOCA 174 101
GWSRYIPEGMQCS OPSD_ZEUFA 174 101
Internet
Start PRIN... Micros... PRINT... http://... Motif3D 9:50 µµ



Για να δείτε την πολλαπλή αντιστοίχηση της opsin, επιλέξτε το “view alignment”.



Επιλέγοντας το “view structure” εμφανίζεται η δομή της OPSIN, όπου μπορείτε να σημαδέψετε τα μοτίβα.



Σύνθετες βάσεις δεδομένων αλληλουχιών πρωτεϊνών

Σε μία σύνθετη βάση δεδομένων συγχωνεύονται διάφορες πρωτογενείς πηγές δεδομένων, καθιστώντας έτσι πιο αποδοτική τη διαδικασία αναζήτησης σειρών. Η βάση αυτή είναι μη πλεονάζουσα (non-redundant – τα δεδομένα δεν επαναλαμβάνονται) και επομένως η διαδικασία αναζήτησης μίας σειράς δε χρειάζεται να πραγματοποιηθεί περισσότερες από μία φορά.
Για τη δημιουργία των σύνθετων βάσεων δεδομένων χρησιμοποιούνται διάφορες στρατηγικές ανάλογα με τις πρωτογενείς πηγές δεδομένων που έχουν επιλεγεί σε κάθε περίπτωση και ανάλογα με τα κριτήρια που χρησιμοποιούνται για τη συγχώνευση των πηγών αυτών. Παραδείγματος χάριν, μία στρατηγική είναι η βάση δεδομένων να μην περιέχει πανομοιότυπες αλληλουχίες και οι αλληλουχίες να διαφέρουν μόνο κατά ένα αμινοξύ (residue – στην περίπτωση αυτή η βάση θα είναι πράγματι μην πλεονάζουσα – non-redundant).
Οι κύριες σύνθετες βάσεις δεδομένων είναι οι ακόλουθες:

	NRDB	OWL	MIPSX	SP+TrEMBL
Πρωτογενείς πηγές δεδομένων	PDB	SWISS-PROT	PIR	SWISS-PROT
	SWISS-PROT	PIR	MIPSOwn	TrEMBL
	PIR	GenBank	MIPSTrn	
	GenPept	NRL-3D		MIPSH
	SWISS-PROTupdate		PIRMOD	
	GenPeptupdate		NRL-3D	
			SWISS-PROT	
			EMTrans	
			GBTrans	
			Kabat	
			PseqIP	

NRDB

Η NRDB (Non-Redundant Database) δημιουργήθηκε στο NCBI. Πρόκειται για μία βάση δεδομένων με εκτενή περιγραφή (έχει όλο το annotation της πρωτεΐνης) η οποία ενημερώνεται συνεχώς. Για τη δημιουργία της αφαιρέθηκαν από τις πρωταρχικές πηγές δεδομένων μόνο τα πανομοιότυπα αντίτυπα των σειρών (επομένως η NRDB είναι μη-πανομοιότυπη, όχι όμως και μη-πλεονάζουσα). Αυτό σημαίνει ότι η βάση περιλαμβάνει πολλά αντίτυπα της ίδιας πρωτεΐνης, τα οποία προκύπτουν ως αποτέλεσμα είτε πολυμορφισμών είτε ελάχιστων λαθών κατά την αλληλούχηση (sequencing). Επιπλέον, λανθασμένες σειρές οι οποίες είχαν διορθωθεί στην SWISS-PROT τώρα επαναλαμβάνονται κατά τη διαδικασία της μετάφρασης από το DNA.

OWL

Η OWL δημιουργήθηκε στο Πανεπιστήμιο του Leeds και είναι μία μη-πλεονάζουσα βάση δεδομένων. Οι πρωτογενείς πηγές δεδομένων της OWL έχουν ιεραρχηθεί και τους έχει αποδοθεί σειρά προτεραιότητας με βάση το επίπεδο περιγραφής (annotation) των πρωτεϊνών, καθώς και την αξιολόγηση των αλληλουχιών που περιλαμβάνουν. Η SWISS-PROT είναι η βάση με την υψηλότερη προτεραιότητα, επομένως οι υπόλοιπες πηγές συγκρίνονται με αυτή κατά τη διαδικασία της

συγχώνευσης. Δεν περιλαμβάνει πανομοιότυπες αλληλουχίες, ούτε αλληλουχίες οι οποίες διαφέρουν μόνο κατά ένα αμινοξύ.

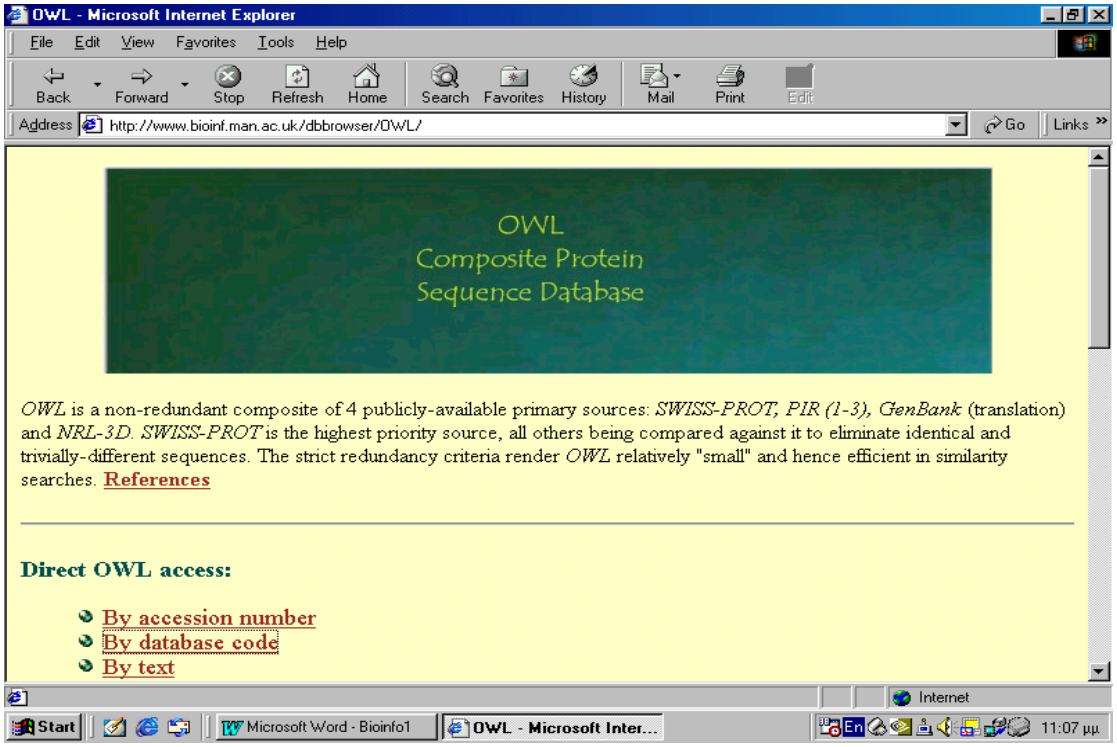
MISPX
Η MIPSX δημιουργήθηκε στο Max-Planck Institut στο Martinsried. Στις πρωτογενείς πηγές της αποδίδεται προτεραιότητα βάσει της σειράς τους όπως αυτή παρουσιάζεται στον άνωθεν πίνακα. Σε περίπτωση που βρεθούν πανομοιότυπες αλληλουχίες είτε μέσα σε κάποια βάση, ή ανάμεσα σε δύο ή περισσότερες βάσεις, τότε οι αλληλουχίες αυτές διαγράφονται αφήνοντας μοναδικά αντίτυπά τους στην MISPX. Επίσης διαγράφονται και οι αλληλουχίες οι οποίες είναι εξ' ολοκλήρου ενσωματωμένες μέσα σε άλλες αλληλουχίες (υπό-αλληλουχίες – subsequences).

SWISS-PROT+TrEMBL
Η SWISS-PROT+TrEMBL δημιουργήθηκε στο EBI και είναι μία βάση δεδομένων εκτενούς περιγραφής και μη-πλεονάζουσα. Εκτιμάται ότι λιγότερο από 30% των δεδομένων του ολικού συνόλου των δεδομένων της SWISS-PROT και της TrEMBL ήταν μη-μοναδικά.

ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

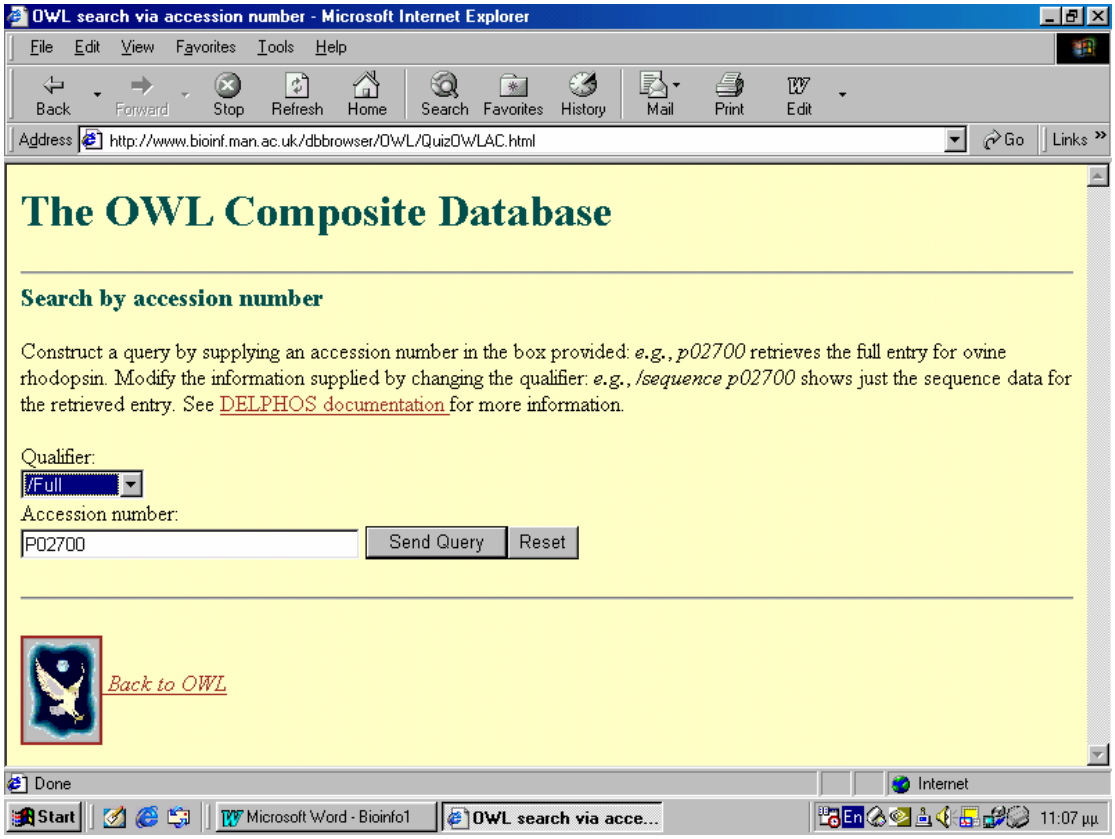
Αναζήτηση (query) χρησιμοποιώντας την OWL

Για πρόσβαση στην OWL ανοίξτε τον Internet Explorer και πληκτρολογήστε την ακόλουθη διεύθυνση:
www.bioinf.man.ac.uk/dbbrowser/OWL/. Στη συνέχεια, για την ανάκτηση πληροφοριών σχετικές με την πρωτεΐνη OPSD_SHEEP με τη χρήση του AC

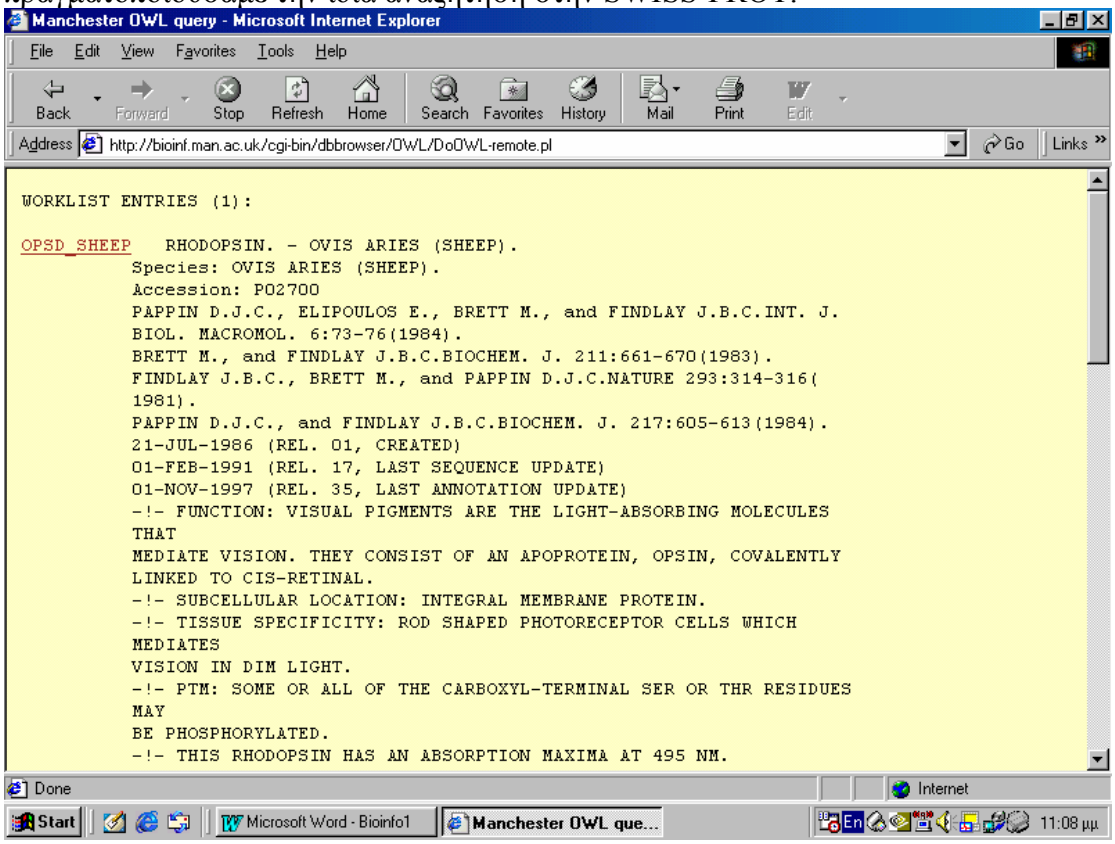


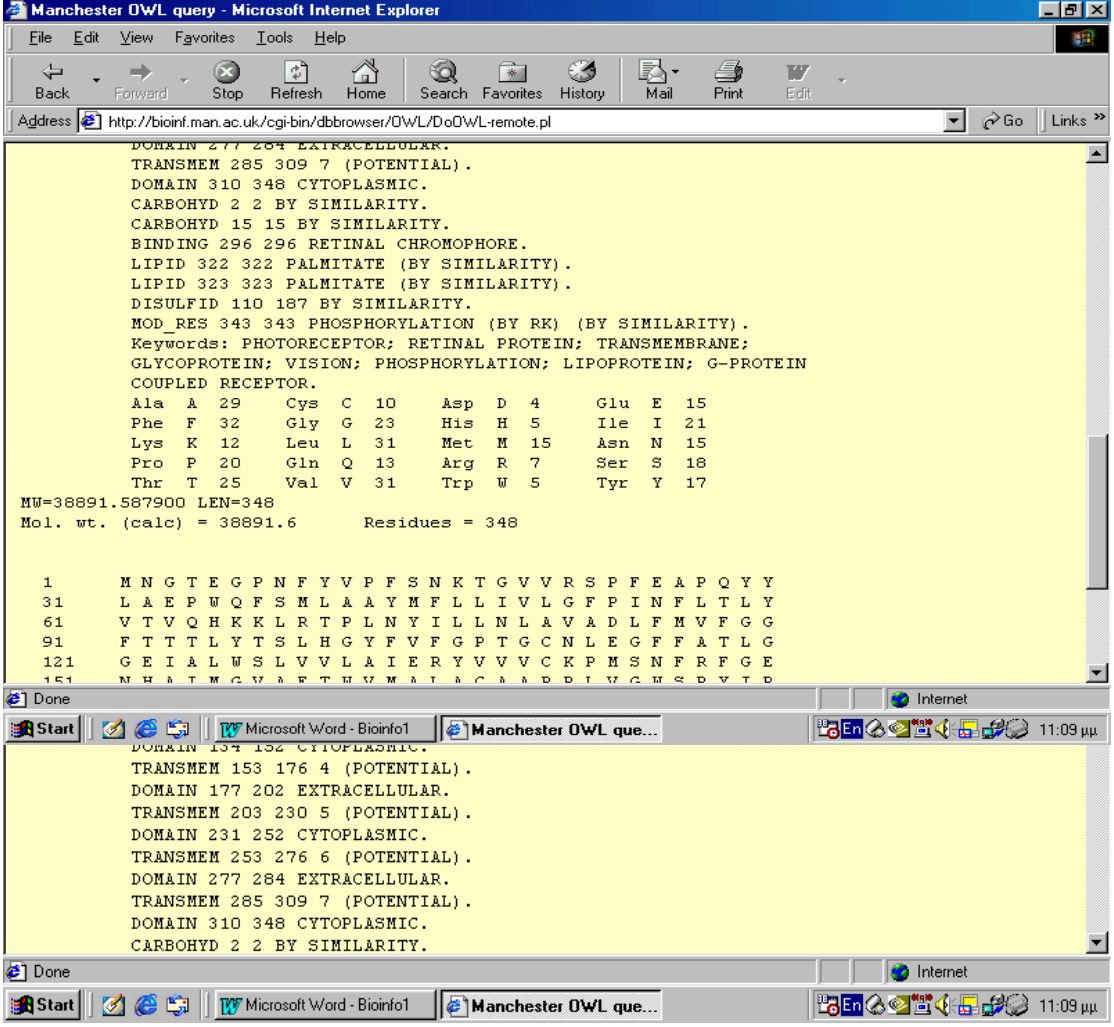
(ACcession number – αριθμού πρόσβασης) για την SWISS-PROT, πατήστε την 'By accession number' επιλογή.

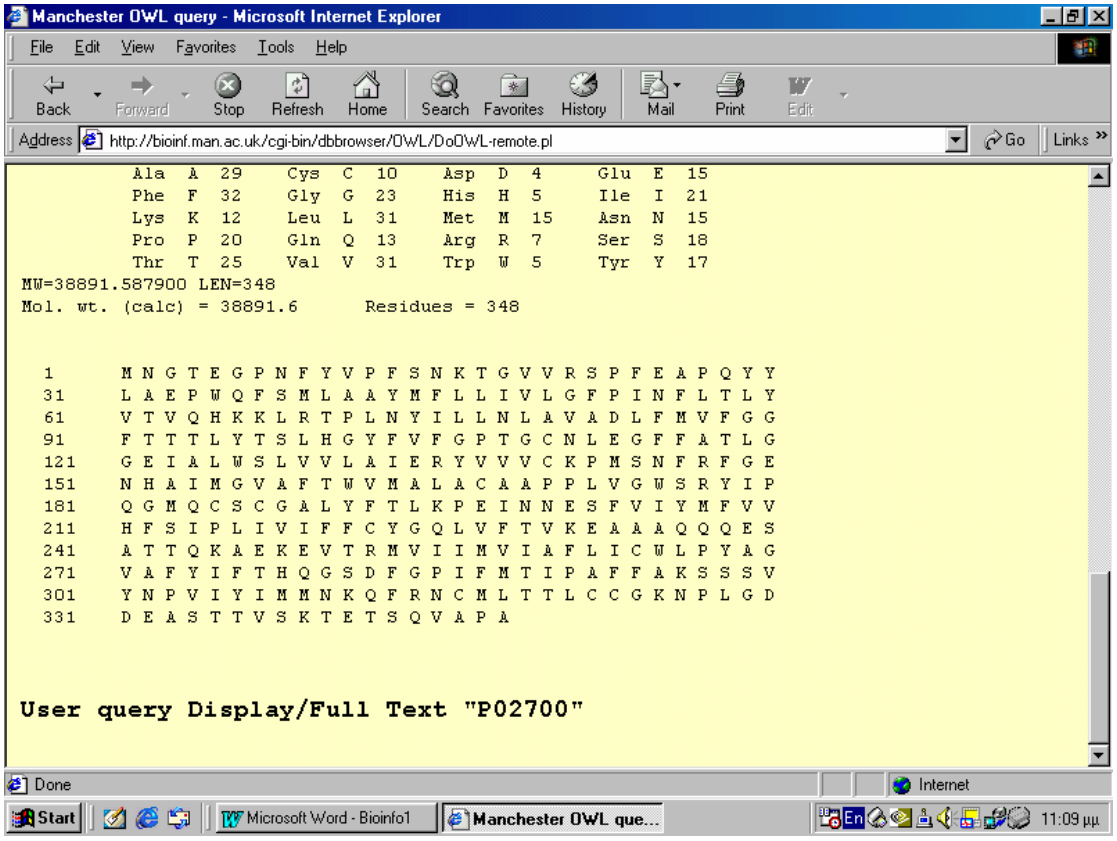
Στην επόμενη σελίδα που θα ανοίξει, πληκτρολογήστε τον αριθμό πρόσβασης (AC) της πρωτεΐνης, ο οποίος είναι P02700, και εν συνεχεία πατήστε 'Send Query'.



Η σελίδα που θα ανοίξει περιλαμβάνει όλα τα στοιχεία που σχετίζονται με την πρωτεΐνη OPSD_SHEEP. Τα στοιχεία αυτά είναι τα ίδια που θα παίρναμε εάν πραγματοποιούσαμε την ίδια αναζήτηση στην SWISS-PROT.



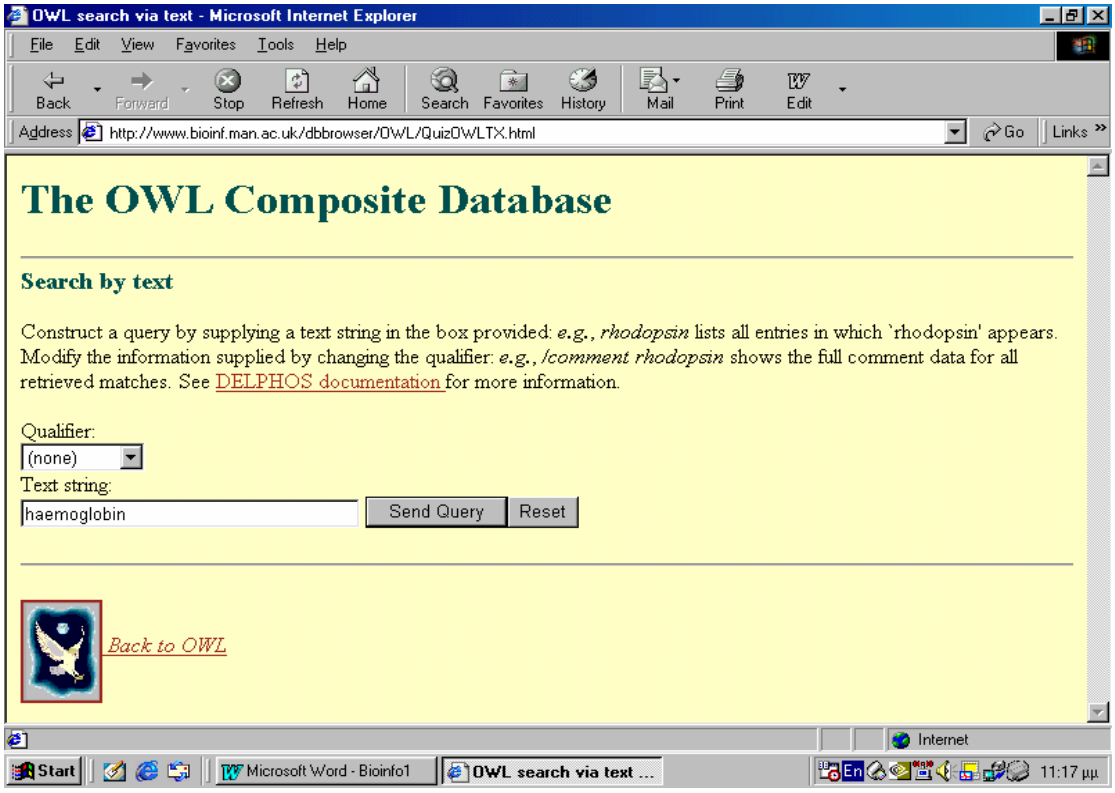




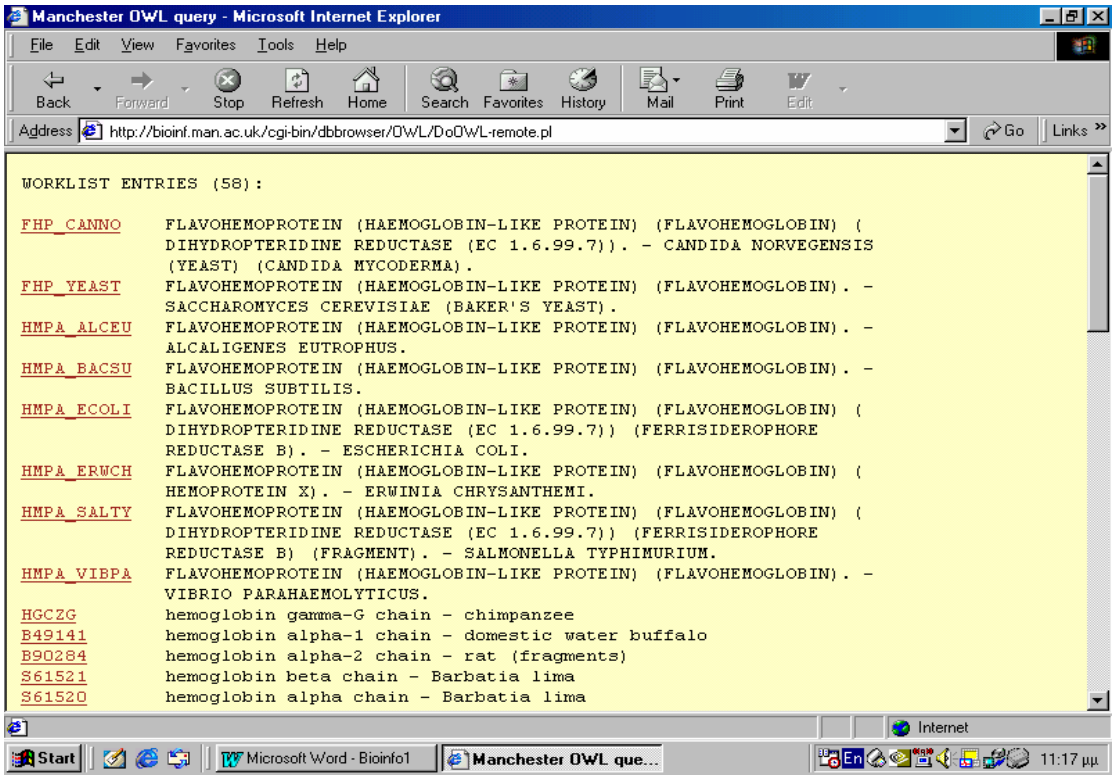
Η επιλογή 'By database code' μπορεί να χρησιμοποιηθεί σε περίπτωση που θέλουμε να ανακτήσουμε πληροφορίες σχετικές με μία πρωτεΐνη χρησιμοποιώντας τον κωδικό που έχει η πρωτεΐνη αυτή στη βάση δεδομένων.

Έτσι, στην περίπτωση του παραδείγματός μας, πατώντας στην επιλογή ‘By database code’, εισάγοντας στην επόμενη σελίδα τον κωδικό της πρωτεΐνης που είναι OPSD_SHEEP και πατώντας την επιλογή ‘Send Query’ μπορούμε πάλι να πάρουμε όλα τα στοιχεία για την πρωτεΐνη OPSD_SHEEP.

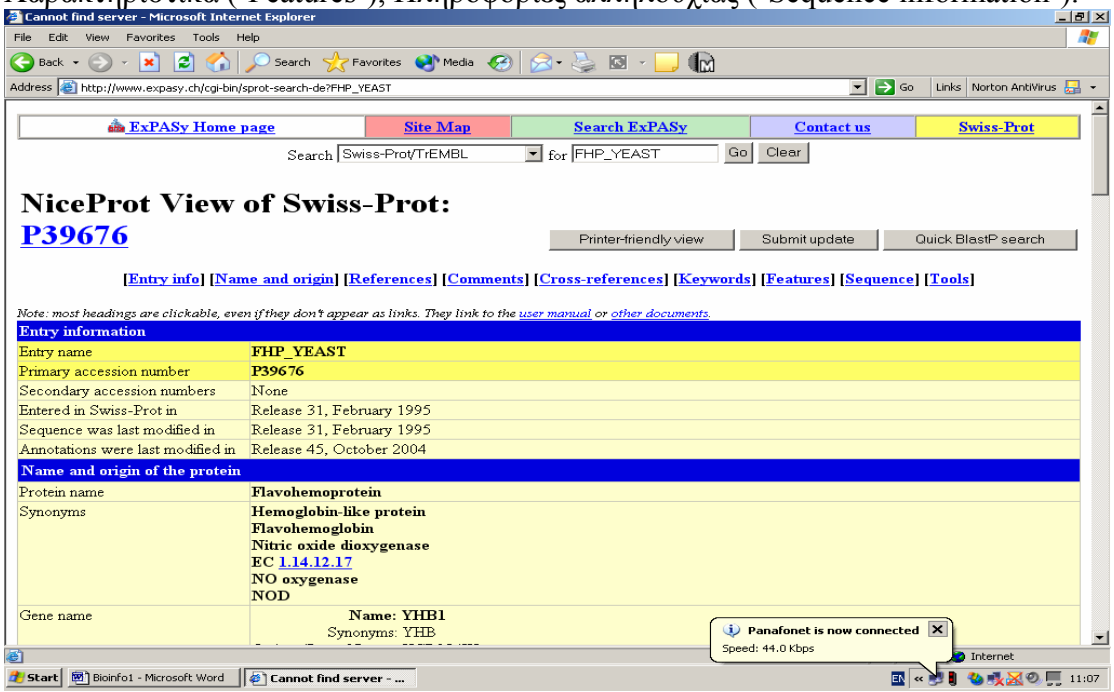
Χρησιμοποιώντας την OWL μπορούμε να πραγματοποιήσουμε και γενική αναζήτηση για μία πρωτεΐνη, π.χ. την Haemoglobin. Σε αυτήν την περίπτωση επιστρέψτε στην πρώτη σελίδα του OWL και πατήστε την επιλογή 'By text'. Πληκτρολογήστε 'haemoglobin' στο κενό πεδίο του 'Text string' και πατήστε 'Send Query'.

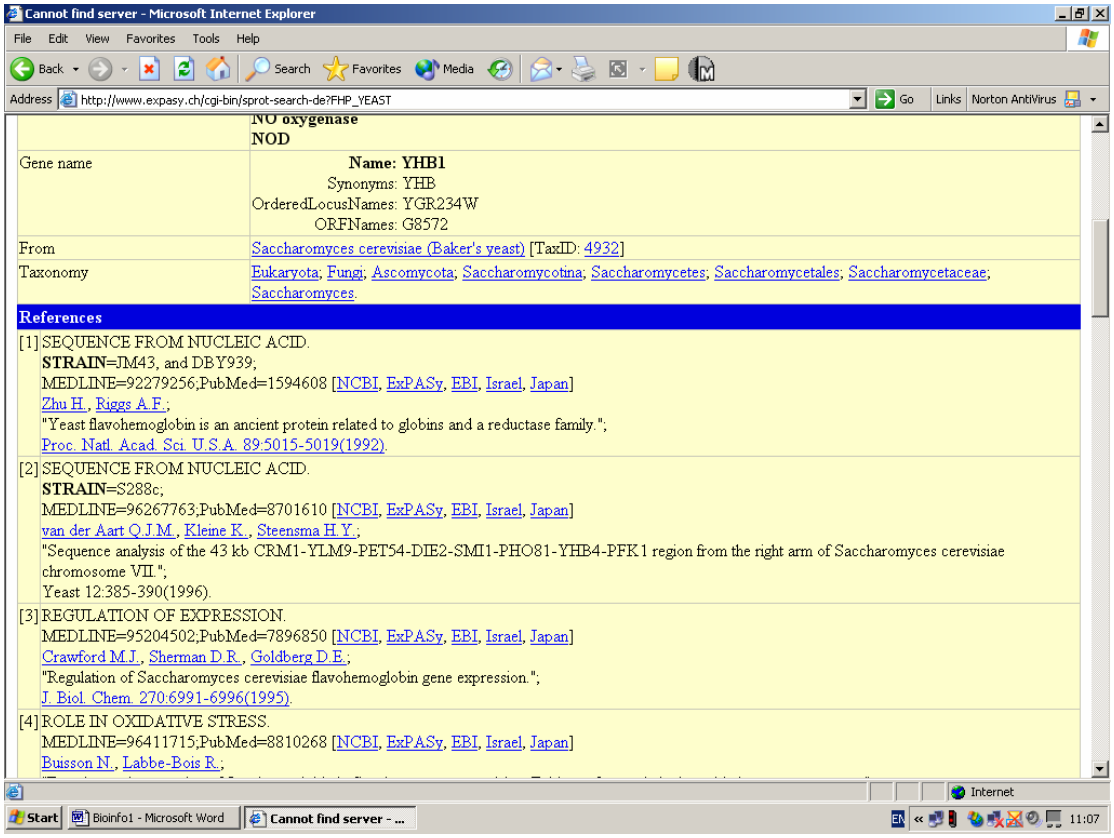


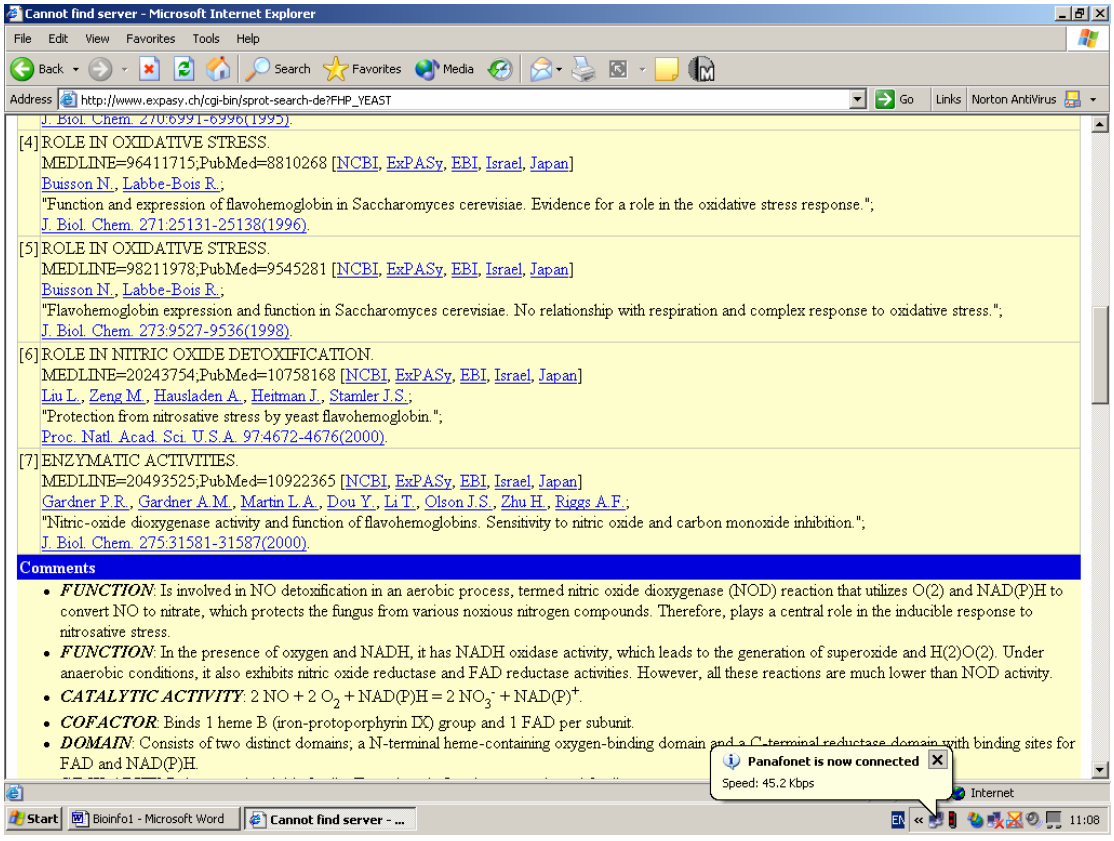
Η σελίδα που θα ανοίξει περιλαμβάνει όλες τις εγγραφές της βάσης για την 'haemoglobin'. Για κάθε μία από τις εγγραφές αυτές μπορείτε να δείτε τα σχετικά αναλυτικά στοιχεία πατώντας πάνω στο όνομά τους. Για παράδειγμα, για τα στοιχεία την εγγραφή FHP_YEAST πατήστε στο 'FHP_YEAST'.

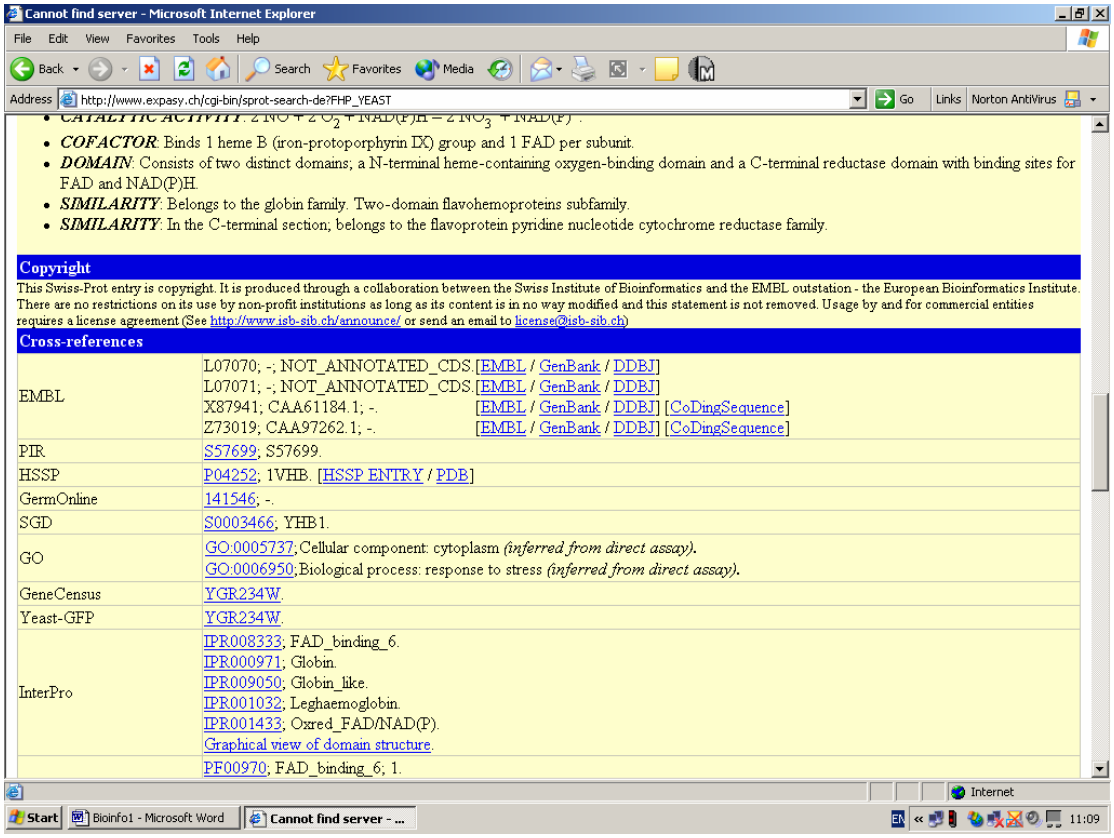


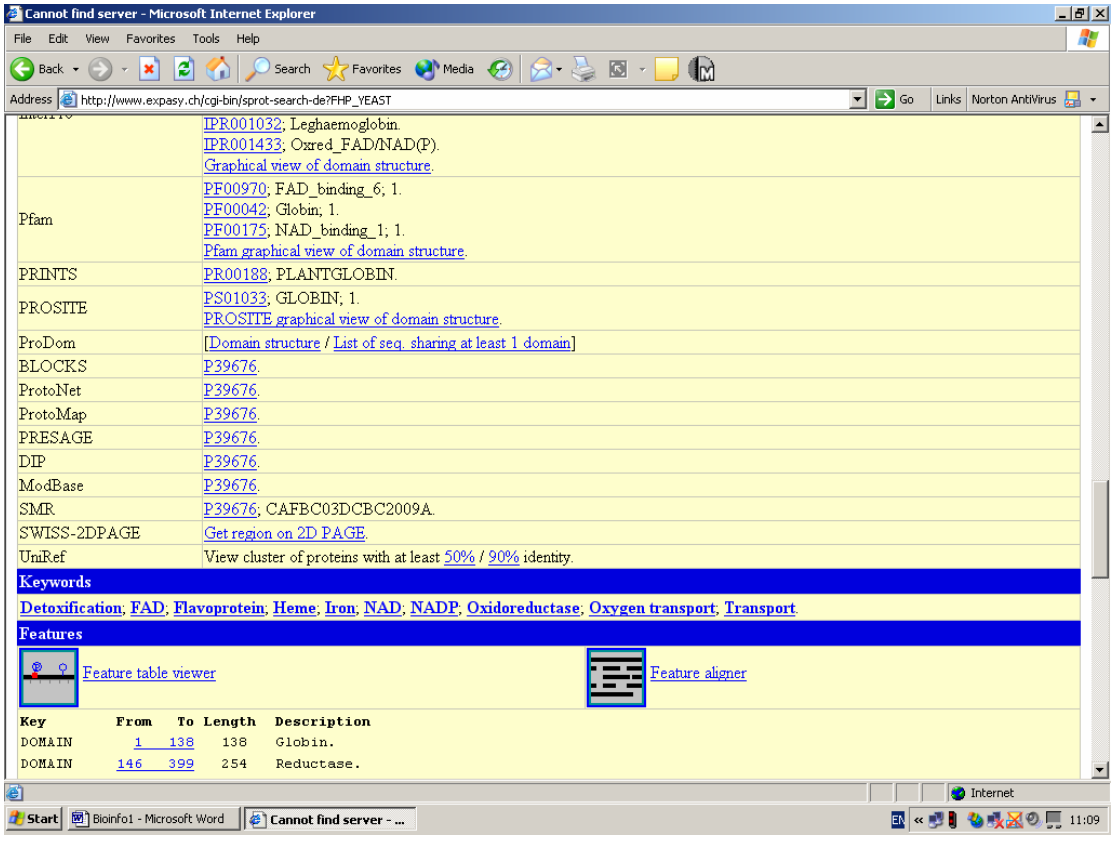
Τότε τα δεδομένα για την FHP_YEAST παρουσιάζονται σε μία νέα σελίδα που ανοίγει από τη βάση δεδομένων SWISS-PROT. Η σελίδα αυτή περιλαμβάνει τις ακόλουθες ενότητες: Πληροφορίες εγγραφής (‘Entry information’), Όνομα και προέλευση της πρωτεΐνης (‘Name and origin of the protein’), Βιβλιογραφία / Αναφορές (‘References’), Σχόλια (‘Comments’), Πνευματικά δικαιώματα (‘Copyright’), Παραπομπές (‘Cross-references’), Λέξη-κλειδί (‘Keyword’), Χαρακτηριστικά (‘Features’), Πληροφορίες αλληλουχίας (‘Sequence information’).

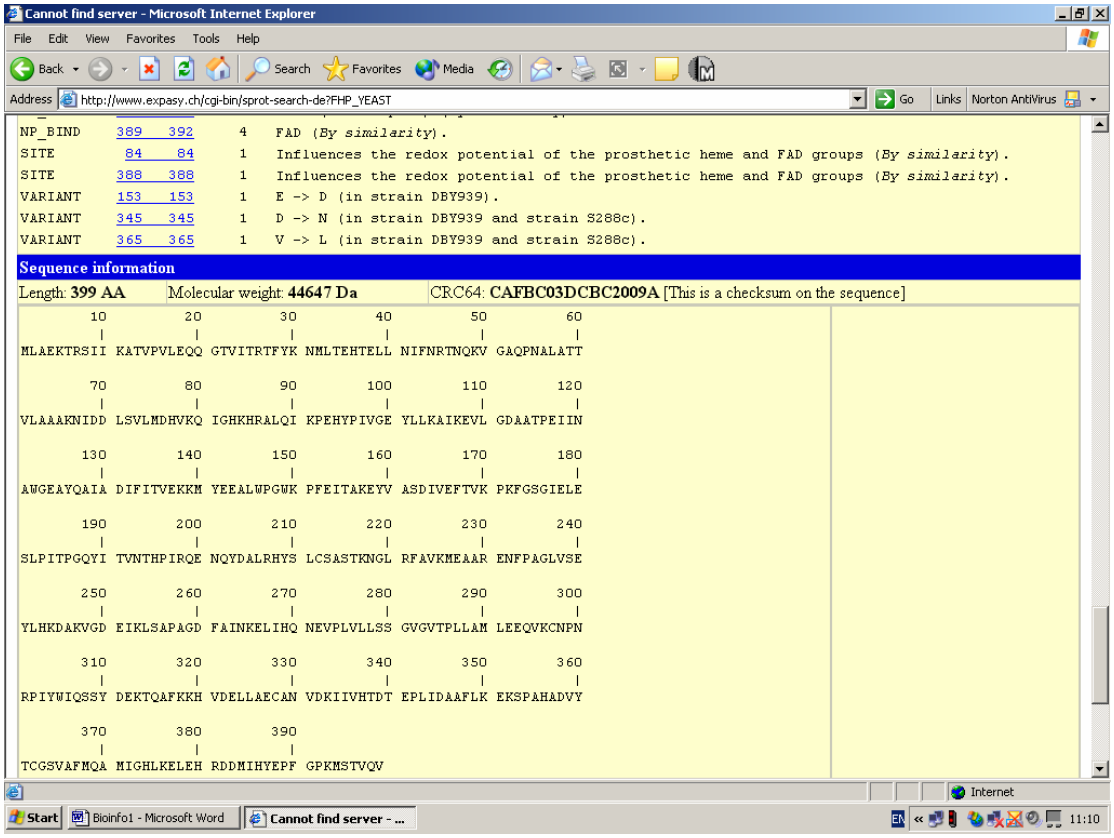



Cannot find server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.expasy.ch/cgi-bin/sprot-search-de?FHP_YEAST
Go Links Norton AntiVirus
NO oxygenase
NOD
Gene name
Name: YHB1
Synonyms: YHB
OrderedLocusNames: YGR234W
ORFNames: G8572
From
Saccharomyces cerevisiae (Baker's yeast) [TaxID: 4932]
Taxonomy
Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.
References
[1] SEQUENCE FROM NUCLEIC ACID.
STRAIN=JM43, and DBY939;
MEDLINE=92279256;PubMed=1594608 [NCBI, ExPASy, EBI, Israel, Japan]
Zhu H., Riggs A.F.;
"Yeast flavohemoglobin is an ancient protein related to globins and a reductase family.";
Proc. Natl. Acad. Sci. U.S.A. 89:5015-5019(1992).
[2] SEQUENCE FROM NUCLEIC ACID.
STRAIN=S288c;
MEDLINE=96267763;PubMed=8701610 [NCBI, ExPASy, EBI, Israel, Japan]
van der Aart Q.J.M., Kleine K., Steensma H.Y.;
"Sequence analysis of the 43 kb CRM1-YLM9-PET54-DIE2-SMI1-PHO81-YHB4-PFK1 region from the right arm of Saccharomyces cerevisiae chromosome VII.";
Yeast 12:385-390(1996).
[3] REGULATION OF EXPRESSION.
MEDLINE=95204502;PubMed=7896850 [NCBI, ExPASy, EBI, Israel, Japan]
Crawford M.J., Sherman D.R., Goldberg D.E.;
"Regulation of Saccharomyces cerevisiae flavohemoglobin gene expression.";
J. Biol. Chem. 270:6991-6996(1995).
[4] ROLE IN OXIDATIVE STRESS.
MEDLINE=96411715;PubMed=8810268 [NCBI, ExPASy, EBI, Israel, Japan]
Buisson N., Labbe-Bois R.;
Internet
Start Bioinfo1 - Microsoft Word Cannot find server - ...
11:07

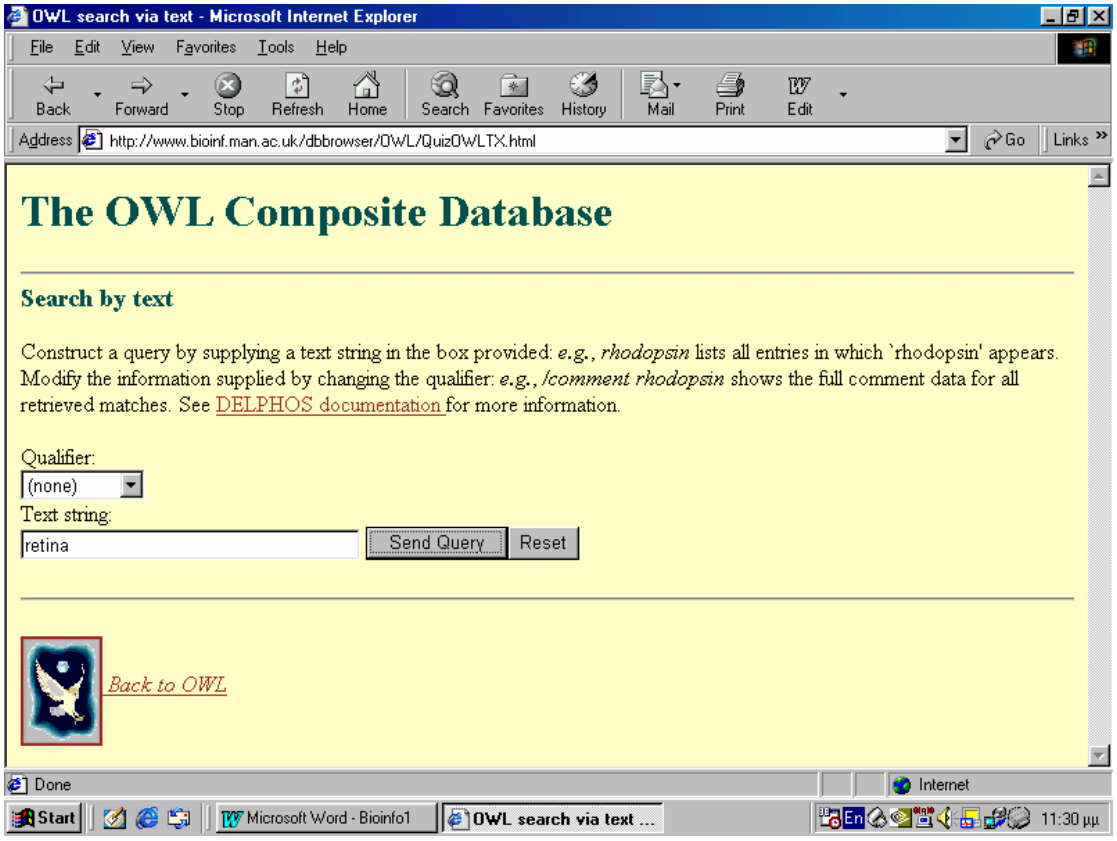

Cannot find server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.expasy.ch/cgi-bin/sprot-search-de?FHP_YEAST
Go Links Norton AntiVirus
J. Biol. Chem. 270:6991-6996(1995).
[4] ROLE IN OXIDATIVE STRESS.
MEDLINE=96411715;PubMed=8810268 [NCBI, ExPASy, EBI, Israel, Japan]
Buisson N., Labbe-Bois R.;
"Function and expression of flavohemoglobin in Saccharomyces cerevisiae. Evidence for a role in the oxidative stress response.";
J. Biol. Chem. 271:25131-25138(1996).
[5] ROLE IN OXIDATIVE STRESS.
MEDLINE=98211978;PubMed=9545281 [NCBI, ExPASy, EBI, Israel, Japan]
Buisson N., Labbe-Bois R.;
"Flavohemoglobin expression and function in Saccharomyces cerevisiae. No relationship with respiration and complex response to oxidative stress.";
J. Biol. Chem. 273:9527-9536(1998).
[6] ROLE IN NITRIC OXIDE DETOXIFICATION.
MEDLINE=20243754;PubMed=10758168 [NCBI, ExPASy, EBI, Israel, Japan]
Liu L., Zeng M., Hausladen A., Heitman J., Stamler J.S.;
"Protection from nitrosative stress by yeast flavohemoglobin.";
Proc. Natl. Acad. Sci. U.S.A. 97:4672-4676(2000).
[7] ENZYMATIC ACTIVITIES.
MEDLINE=20493525;PubMed=10922365 [NCBI, ExPASy, EBI, Israel, Japan]
Gardner P.R., Gardner A.M., Martin L.A., Dou Y., Li T., Olson J.S., Zhu H., Riggs A.F.;
"Nitric-oxide dioxygenase activity and function of flavohemoglobins. Sensitivity to nitric oxide and carbon monoxide inhibition.";
J. Biol. Chem. 275:31581-31587(2000).
Comments
FUNCTION: Is involved in NO detoxification in an aerobic process, termed nitric oxide dioxygenase (NOD) reaction that utilizes O(2) and NAD(P)H to convert NO to nitrate, which protects the fungus from various noxious nitrogen compounds. Therefore, plays a central role in the inducible response to nitrosative stress.
FUNCTION: In the presence of oxygen and NADH, it has NADH oxidase activity, which leads to the generation of superoxide and H(2)O(2). Under anaerobic conditions, it also exhibits nitric oxide reductase and FAD reductase activities. However, all these reactions are much lower than NOD activity.
CATALYTIC ACTIVITY: 2 NO + 2 O2 + NAD(P)H = 2 NO3- + NAD(P)+.
COFACTOR: Binds 1 heme B (iron-protoporphyrin IX) group and 1 FAD per subunit.
DOMAIN: Consists of two distinct domains; a N-terminal heme-containing oxygen-binding domain and a C-terminal reductase domain with binding sites for FAD and NAD(P)H.
Panafonet is now connected
Speed: 45.2 Kbps
Internet
Start Bioinfo1 - Microsoft Word Cannot find server - ...
11:08


Cannot find server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.expasy.ch/cgi-bin/sprot-search-de?FHP_YEAST
Go Links Norton AntiVirus
CATALYTIC ACTIVITY: $2\,NO + 2\,O_2 + NAD(P)H = 2\,NO_3^- + NAD(P)^+$.
COFACTOR: Binds 1 heme B (iron-protoporphyrin IX) group and 1 FAD per subunit.
DOMAIN: Consists of two distinct domains; a N-terminal heme-containing oxygen-binding domain and a C-terminal reductase domain with binding sites for FAD and NAD(P)H.
SIMILARITY: Belongs to the globin family. Two-domain flavohemoproteins subfamily.
SIMILARITY: In the C-terminal section; belongs to the flavoprotein pyridine nucleotide cytochrome reductase family.
Copyright
This Swiss-Prot entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See http://www.isb-sib.ch/announce/ or send an email to license@isb-sib.ch)
Cross-references
EMBL
L07070; -; NOT_ANNOTATED_CDS. [EMBL / GenBank / DDBJ]
L07071; -; NOT_ANNOTATED_CDS. [EMBL / GenBank / DDBJ]
X87941; CAA61184.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
Z73019; CAA97262.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR
S57699; S57699.
HSSP
P04252; 1VHB. [HSSP ENTRY / PDB]
GermOnline
141546; -.
SGD
S0003466; YHB1.
GO
GO:0005737; Cellular component: cytoplasm (inferred from direct assay).
GO:0006950; Biological process: response to stress (inferred from direct assay).
GeneCensus
YGR234W.
Yeast-GFP
YGR234W.
InterPro
IPR008333; FAD_binding_6.
IPR000971; Globin.
IPR009050; Globin_like.
IPR001032; Leghaemoglobin.
IPR001433; Oxred_FAD/NAD(P).
Graphical view of domain structure.
PF00970; FAD_binding_6; 1.
Internet
Start Bioinfo1 - Microsoft Word Cannot find server - ...
11:09

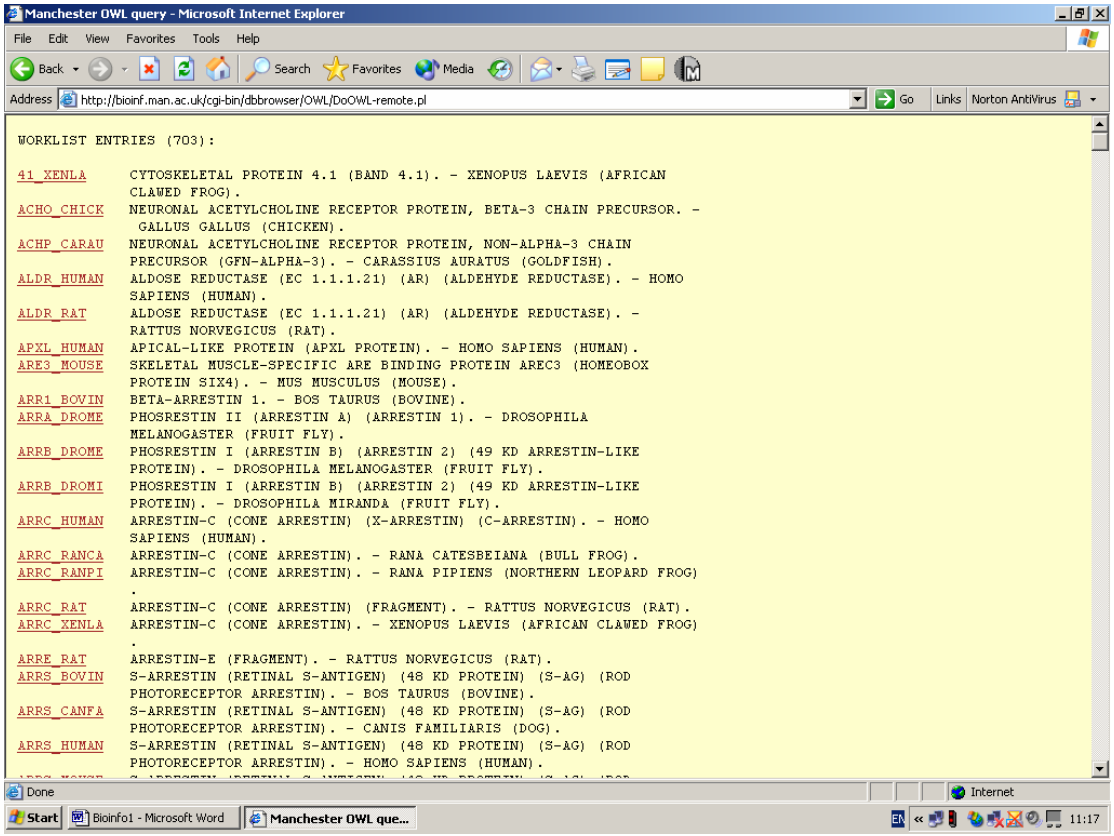

Cannot find server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.expasy.ch/cgi-bin/sprot-search-de?FHP_YEAST
Go Links Norton AntiVirus
IPR001032; Leghaemoglobin.
IPR001433; Oxred_FAD/NAD(P).
Graphical view of domain structure.
Pfam
PF00970; FAD_binding_6; 1.
PF00042; Globin; 1.
PF00175; NAD_binding_1; 1.
Pfam graphical view of domain structure.
PRINTS
PR00188; PLANTGLOBIN.
PROSITE
PS01033; GLOBIN; 1.
PROSITE graphical view of domain structure.
ProDom
[Domain structure / List of seq. sharing at least 1 domain]
BLOCKS
P39676.
ProtoNet
P39676.
ProtoMap
P39676.
PRESAGE
P39676.
DIP
P39676.
ModBase
P39676.
SMR
P39676; CAFBC03DCBC2009A.
SWISS-2DPAGE
Get region on 2D PAGE.
UniRef
View cluster of proteins with at least 50% / 90% identity.
Keywords
Detoxification; FAD; Flavoprotein; Heme; Iron; NAD; NADP; Oxidoreductase; Oxygen transport; Transport.
Features
Feature table viewer
Feature aligner
Key From To Length Description
DOMAIN 1 138 138 Globin.
DOMAIN 146 399 254 Reductase.
Internet
Start Bioinfo1 - Microsoft Word Cannot find server - ...
11:09


Cannot find server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.expasy.ch/cgi-bin/sprot-search-de?FHP_YEAST
Go Links Norton AntiVirus
NP_BIND 389 392 4 FAD (By similarity).
SITE 84 84 1 Influences the redox potential of the prosthetic heme and FAD groups (By similarity).
SITE 388 388 1 Influences the redox potential of the prosthetic heme and FAD groups (By similarity).
VARIANT 153 153 1 E -> D (in strain DBY939).
VARIANT 345 345 1 D -> N (in strain DBY939 and strain S288c).
VARIANT 365 365 1 V -> L (in strain DBY939 and strain S288c).
Sequence information
Length: 399 AA
Molecular weight: 44647 Da
CRC64: CAFBC03DCBC2009A [This is a checksum on the sequence]
10 20 30 40 50 60
MLAEKTRSII KATVPVLEQQ GTVITRTFYK NMLTEHTELL NIFNRTNQKV GAQPNALATT
70 80 90 100 110 120
VLAAAKNIDD LSVLMDHVKQ IGHKHRALQI KPEHYPIVGE YLLKAIKEVL GDAATPEIIN
130 140 150 160 170 180
AWGEAYQAIA DIFITVEKKM YEEALWPGWK PFEITAKEYV ASDIVEFTVK PKFGSGIELE
190 200 210 220 230 240
SLPITPGQYI TVNTHPIRQE NQYDALRHYS LCSASTKNGL RFAVKMEAAR ENFPAGLVSE
250 260 270 280 290 300
YLHKDAKVGD EIKLSAPAGD FAINKELIHQ NEVPLVLLSS GVGVTPLLAM LEEQVKCNPN
310 320 330 340 350 360
RPIYWIQSSY DEKTQAFKKH VDELLAECAN VDKIIVHTDT EPLIDAAFLK EKSPAHADVY
370 380 390
TCGSVAFMQA MIGHLKELEH RDDMIHYEPF GPKMSTVQV
Internet
Start
Bioinfo1 - Microsoft Word
Cannot find server - ...
11:10

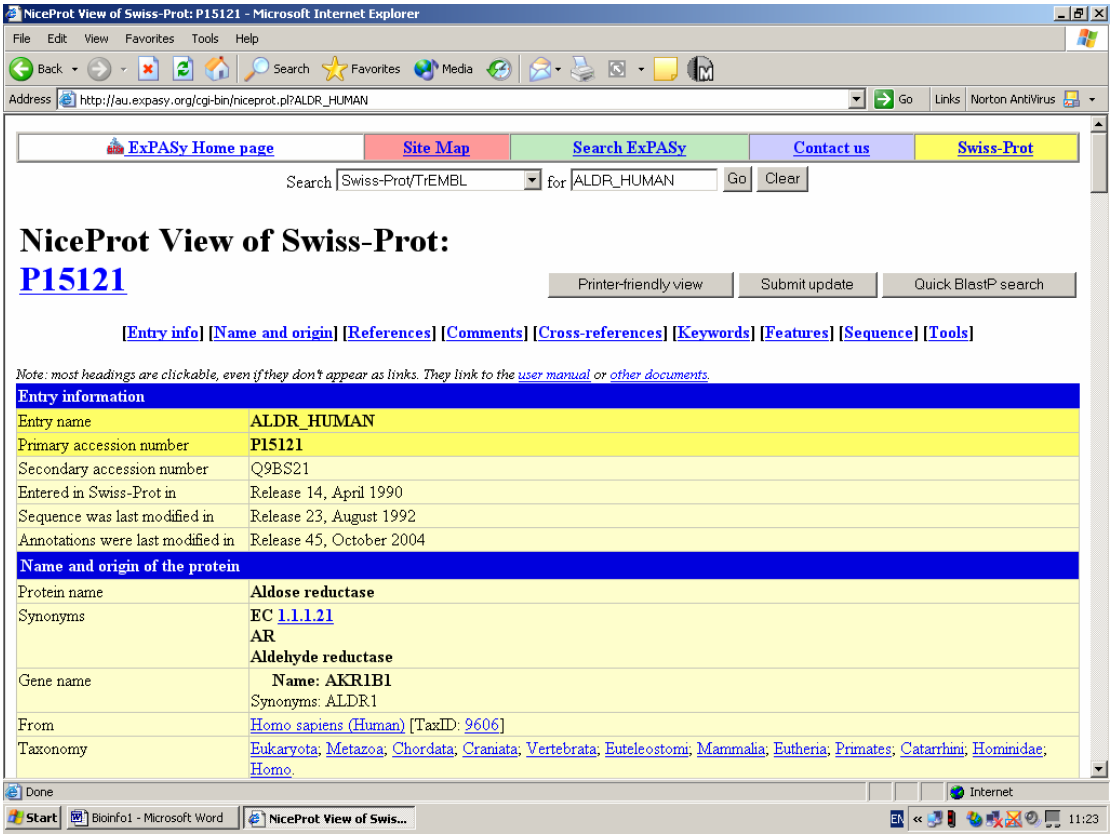
Μπορείτε επίσης να αναζητήσετε πρωτεΐνες σχετικές με το όνομα ‘retina’

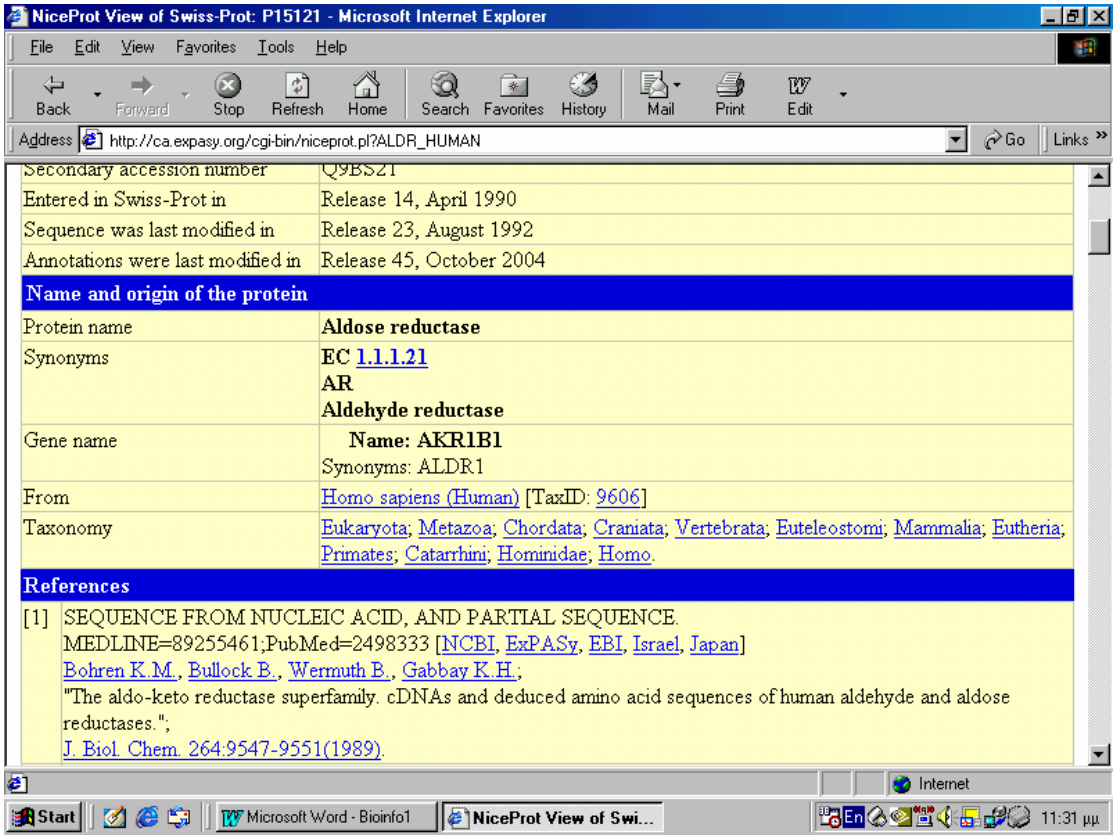


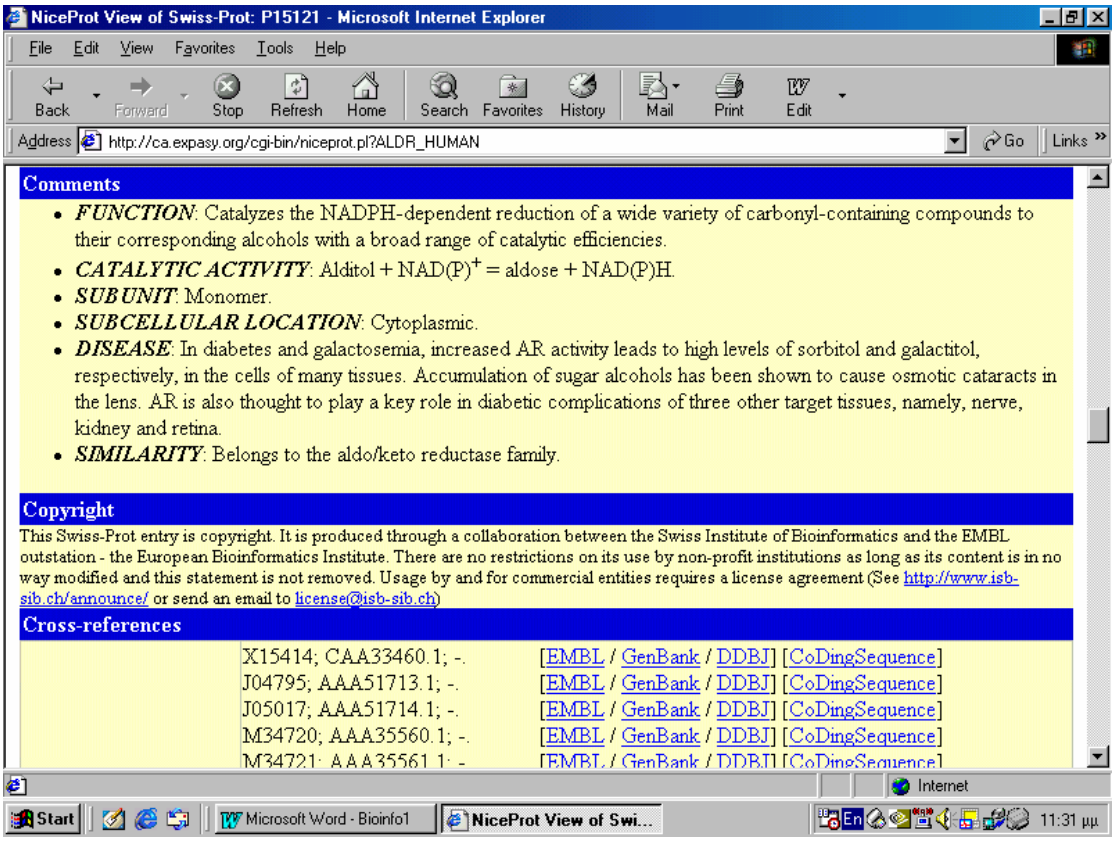
πληκτρολογώντας retina στο κενό πεδίο του Text string και πατώντας το ‘Send query’.
Η σελίδα που θα ανοίξει περιλαμβάνει όλες τις πρωτεΐνες που σχετίζονται με την retina και που βρίσκονται αποθηκευμένες στην OWL. Πατήστε στην ALDR_HUMAN για να δείτε την εγγραφή για αυτήν την πρωτεΐνη.

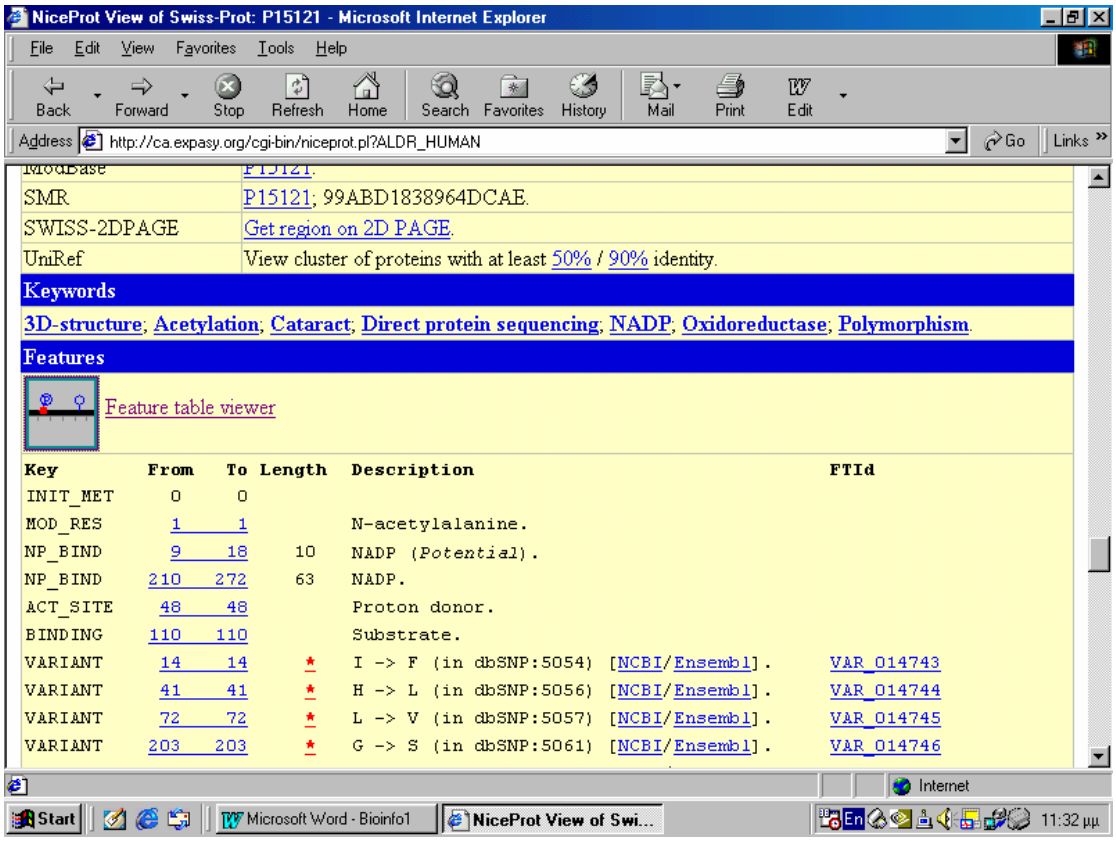


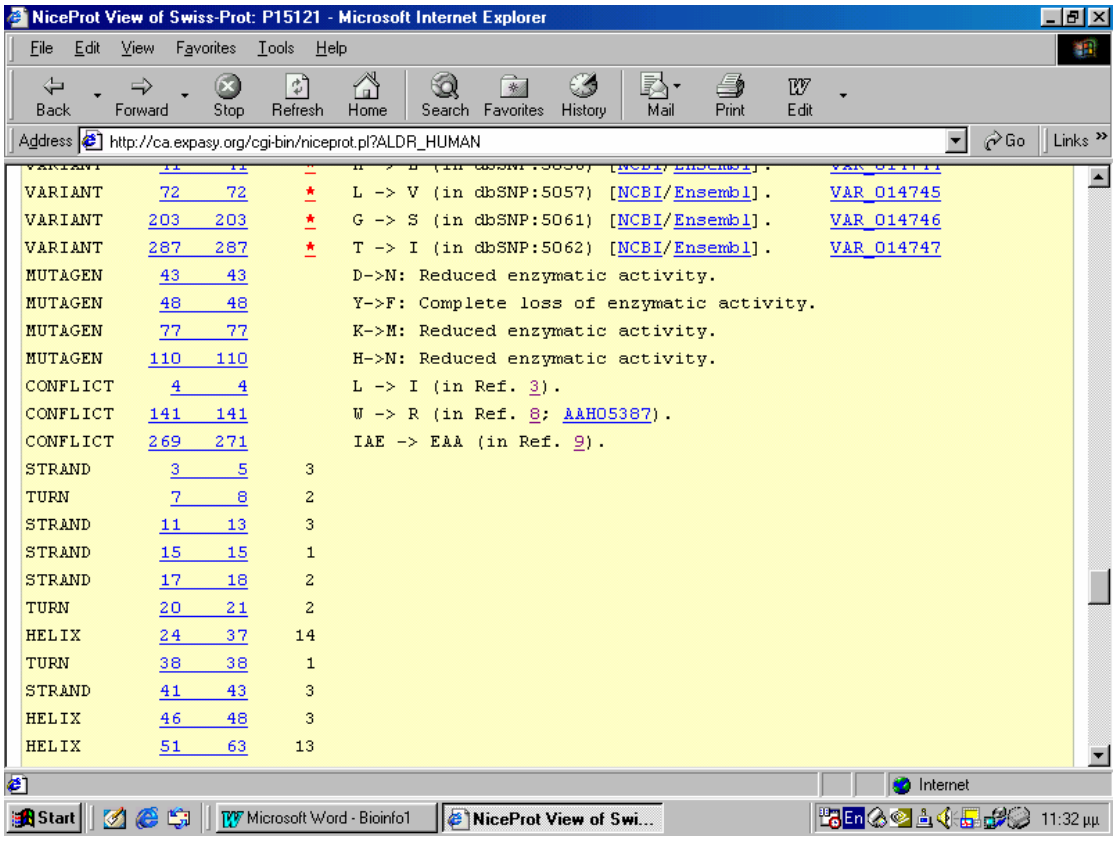
Στη συνέχεια, μέσω της SWISS-PROT μπορείτε να δείτε τη σελίδα με όλες τις πληροφορίες για την ALDR_HUMAN.

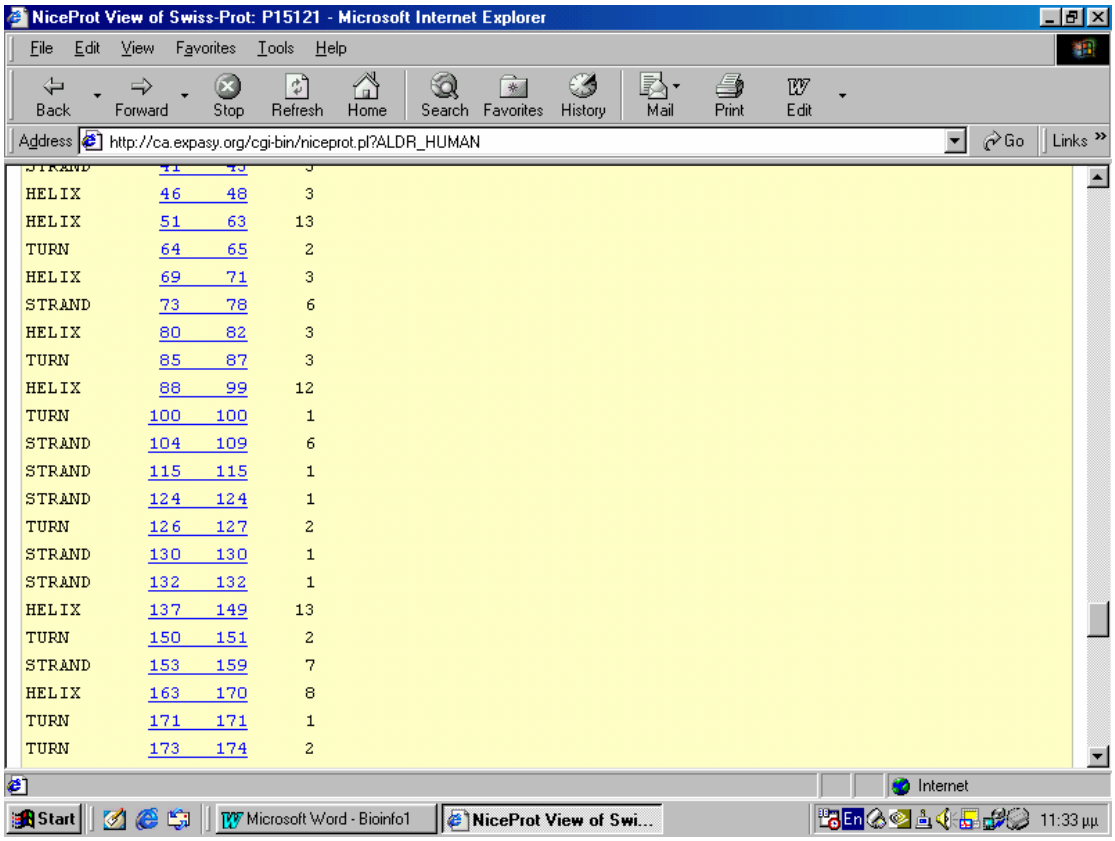

NiceProt View of Swiss-Prot: P15121 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://au.expasy.org/cgi-bin/niceprot.pl?ALDR_HUMAN
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Search Swiss-Prot/TrEMBL for ALDR_HUMAN Go Clear
NiceProt View of Swiss-Prot:
P15121
Printer-friendly view
Submit update
Quick BlastP search
[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]
Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.
Entry information
Entry name ALDR_HUMAN
Primary accession number P15121
Secondary accession number Q9BS21
Entered in Swiss-Prot in Release 14, April 1990
Sequence was last modified in Release 23, August 1992
Annotations were last modified in Release 45, October 2004
Name and origin of the protein
Protein name Aldose reductase
Synonyms EC 1.1.1.21
AR
Aldehyde reductase
Gene name Name: AKR1B1
Synonyms: ALDR1
From Homo sapiens (Human) [TaxID: 9606]
Taxonomy Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
Done
Internet
Start
Bioinfo1 - Microsoft Word
NiceProt View of Swis...
11:23

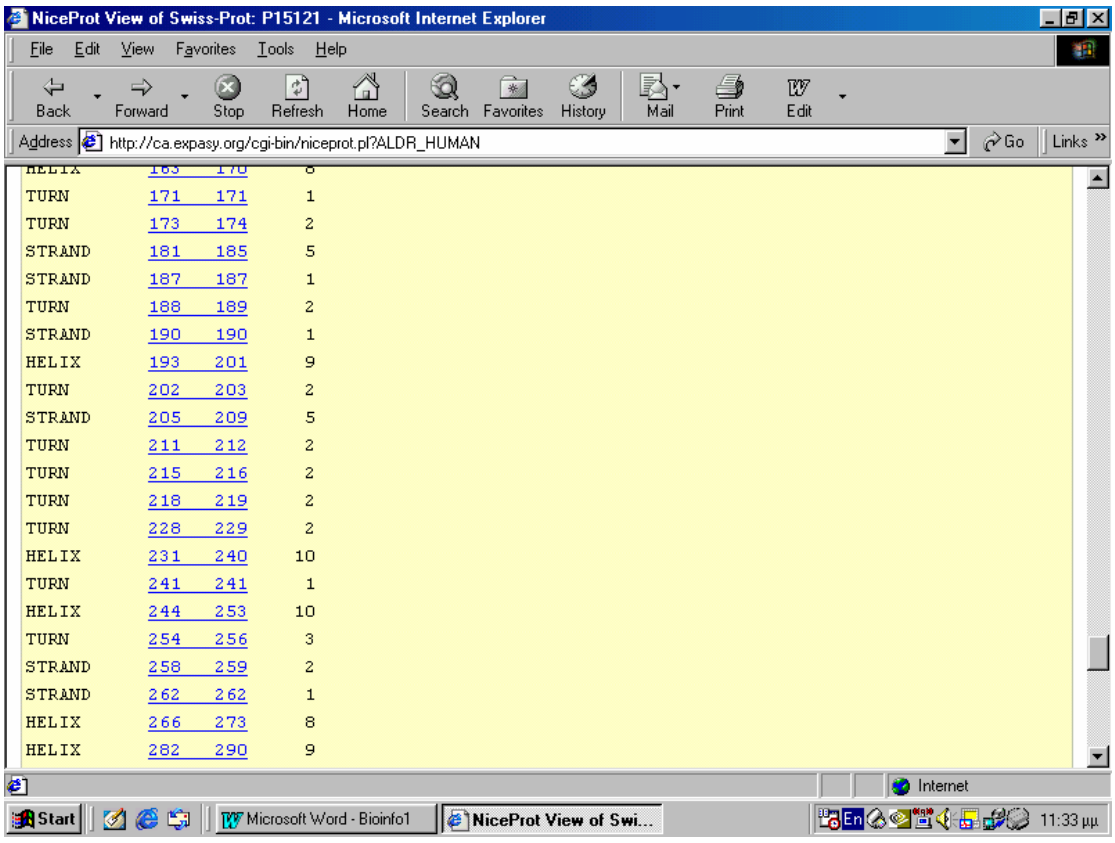

NiceProt View of Swiss-Prot: P15121 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/niceprot.pl?ALDR_HUMAN
Go Links
Secondary accession number Q9BS21
Entered in Swiss-Prot in Release 14, April 1990
Sequence was last modified in Release 23, August 1992
Annotations were last modified in Release 45, October 2004
Name and origin of the protein
Protein name Aldose reductase
Synonyms EC 1.1.1.21
AR
Aldehyde reductase
Gene name Name: AKR1B1
Synonyms: ALDR1
From Homo sapiens (Human) [TaxID: 9606]
Taxonomy Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
References
[1] SEQUENCE FROM NUCLEIC ACID, AND PARTIAL SEQUENCE.
MEDLINE=89255461;PubMed=2498333 [NCBI, ExPASy, EBI, Israel, Japan]
Bohren K.M., Bullock B., Wermuth B., Gabbay K.H.;
"The aldo-keto reductase superfamily. cDNAs and deduced amino acid sequences of human aldehyde and aldose reductases.";
J. Biol. Chem. 264:9547-9551(1989).
Internet
Start
Microsoft Word - Bioinfo1
NiceProt View of Swi...
11:31 μμ


NiceProt View of Swiss-Prot: P15121 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/niceprot.pl?ALDR_HUMAN
Go Links
Comments
FUNCTION: Catalyzes the NADPH-dependent reduction of a wide variety of carbonyl-containing compounds to their corresponding alcohols with a broad range of catalytic efficiencies.
CATALYTIC ACTIVITY: Alditol + NAD(P)+ = aldose + NAD(P)H.
SUBUNIT: Monomer.
SUBCELLULAR LOCATION: Cytoplasmic.
DISEASE: In diabetes and galactosemia, increased AR activity leads to high levels of sorbitol and galactitol, respectively, in the cells of many tissues. Accumulation of sugar alcohols has been shown to cause osmotic cataracts in the lens. AR is also thought to play a key role in diabetic complications of three other target tissues, namely, nerve, kidney and retina.
SIMILARITY: Belongs to the aldo/keto reductase family.
Copyright
This Swiss-Prot entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See http://www.isb-sib.ch/announce/ or send an email to license@isb-sib.ch)
Cross-references
X15414; CAA33460.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
J04795; AAA51713.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
J05017; AAA51714.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
M34720; AAA35560.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
Internet
Start
Microsoft Word - Bioinfo1
NiceProt View of Swi...
11:31 μμ

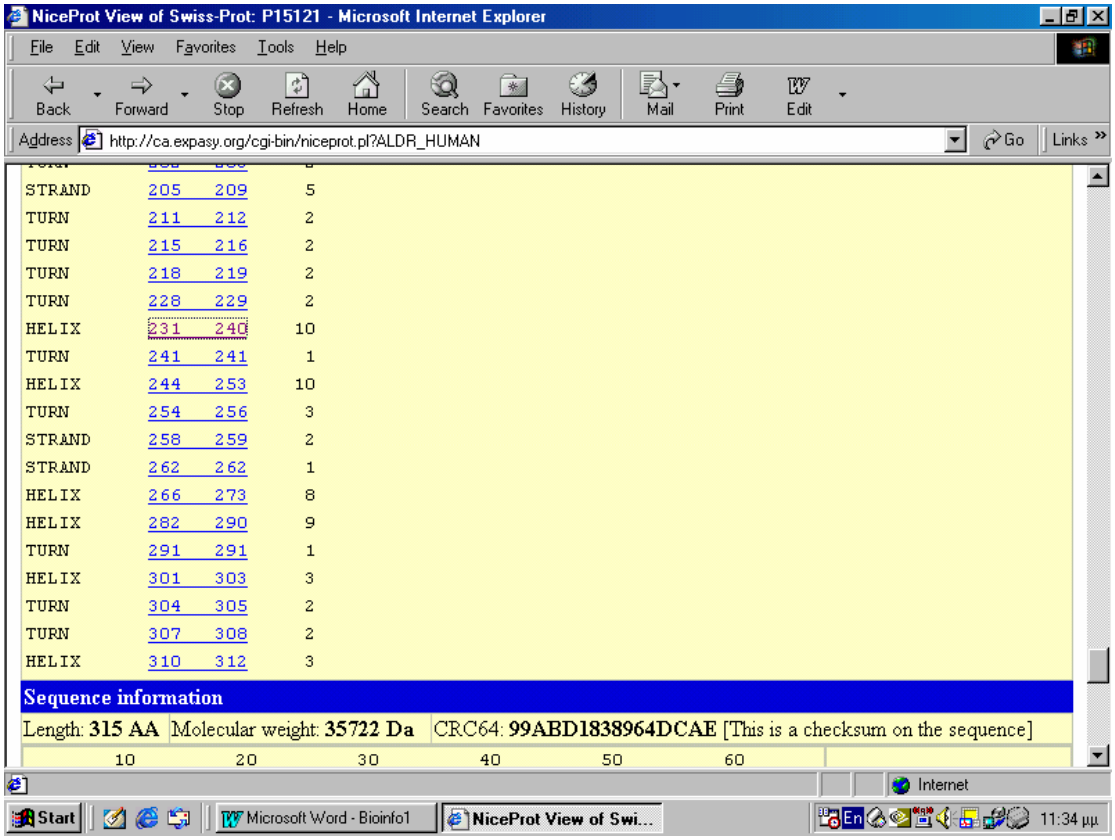

NiceProt View of Swiss-Prot: P15121 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/niceprot.pl?ALDR_HUMAN
Go Links
ModBase P15121.
SMR P15121; 99ABD1838964DCAE.
SWISS-2DPAGE Get region on 2D PAGE.
UniRef View cluster of proteins with at least 50% / 90% identity.
Keywords
3D-structure; Acetylation; Cataract; Direct protein sequencing; NADP; Oxidoreductase; Polymorphism.
Features
Feature table viewer
Key From To Length Description FTId
INIT_MET 0 0
MOD_RES 1 1 N-acetylalanine.
NP_BIND 9 18 10 NADP (Potential).
NP_BIND 210 272 63 NADP.
ACT_SITE 48 48 Proton donor.
BINDING 110 110 Substrate.
VARIANT 14 14 * I -> F (in dbSNP:5054) [NCBI/Ensembl]. VAR_014743
VARIANT 41 41 * H -> L (in dbSNP:5056) [NCBI/Ensembl]. VAR_014744
VARIANT 72 72 * L -> V (in dbSNP:5057) [NCBI/Ensembl]. VAR_014745
VARIANT 203 203 * G -> S (in dbSNP:5061) [NCBI/Ensembl]. VAR_014746
Internet
Start Microsoft Word - Bioinfo1 NiceProt View of Swi... 11:32 μμ

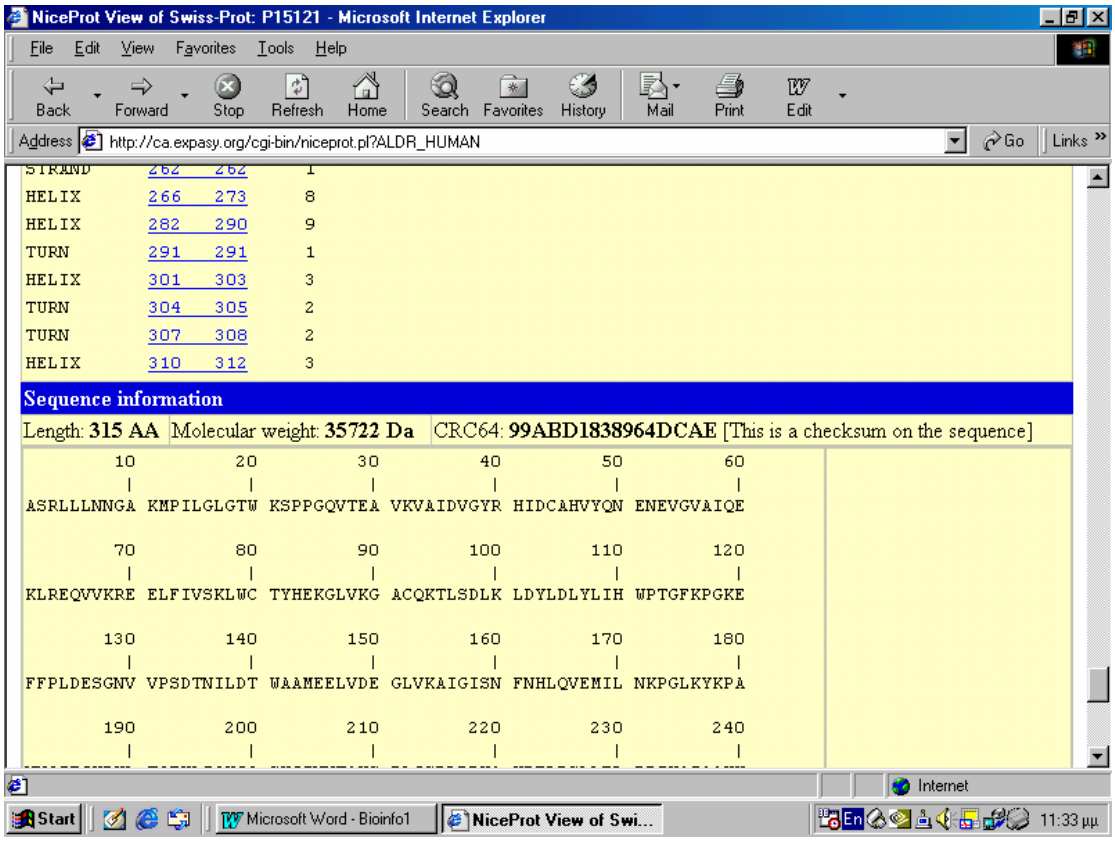

NiceProt View of Swiss-Prot: P15121 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/niceprot.pl?ALDR_HUMAN
Go Links
VARIANT 72 72 * L -> V (in dbSNP:5057) [NCBI/Ensembl]. VAR_014745
VARIANT 203 203 * G -> S (in dbSNP:5061) [NCBI/Ensembl]. VAR_014746
VARIANT 287 287 * T -> I (in dbSNP:5062) [NCBI/Ensembl]. VAR_014747
MUTAGEN 43 43 D->N: Reduced enzymatic activity.
MUTAGEN 48 48 Y->F: Complete loss of enzymatic activity.
MUTAGEN 77 77 K->M: Reduced enzymatic activity.
MUTAGEN 110 110 H->N: Reduced enzymatic activity.
CONFLICT 4 4 L -> I (in Ref. 3).
CONFLICT 141 141 W -> R (in Ref. 8; AAH05387).
CONFLICT 269 271 IAE -> EAA (in Ref. 9).
STRAND 3 5 3
TURN 7 8 2
STRAND 11 13 3
STRAND 15 15 1
STRAND 17 18 2
TURN 20 21 2
HELIX 24 37 14
TURN 38 38 1
STRAND 41 43 3
HELIX 46 48 3
HELIX 51 63 13
Internet
Start Microsoft Word - Bioinfo1 NiceProt View of Swi... 11:32 μμ

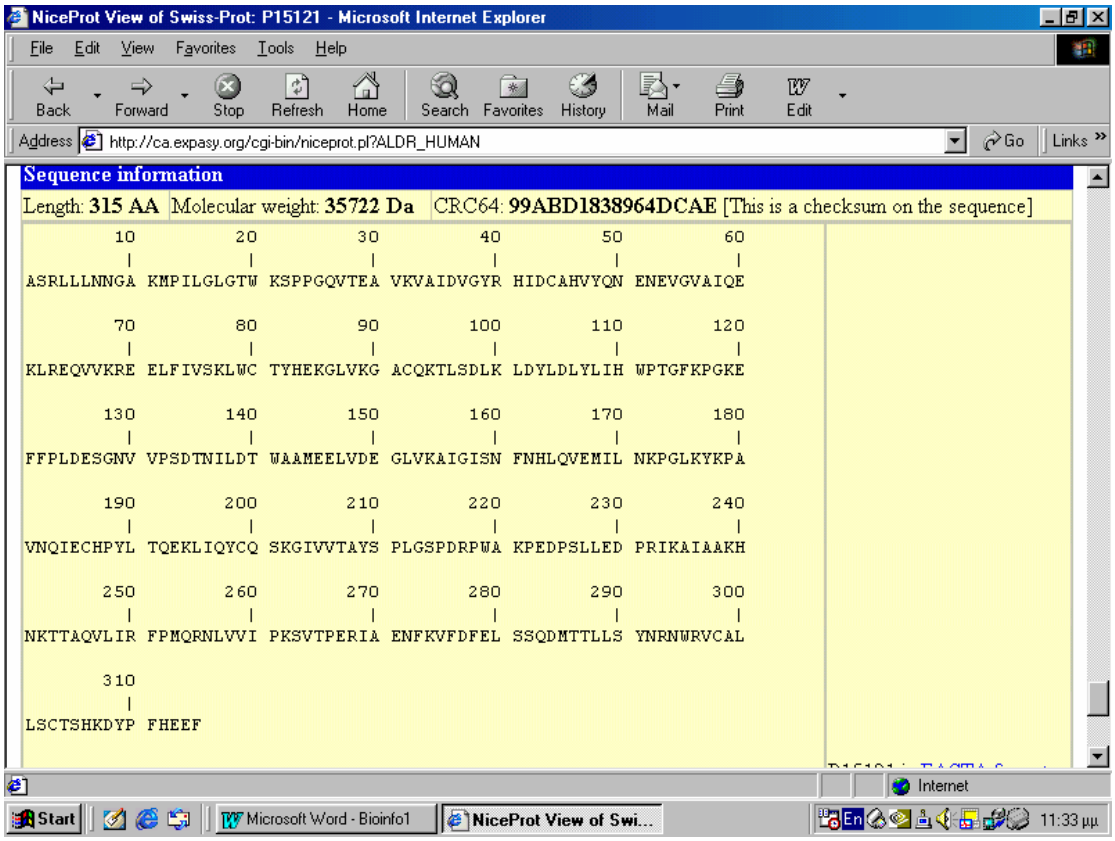

NiceProt View of Swiss-Prot: P15121 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/niceprot.pl?ALDR_HUMAN
Go Links
STRAND 41 43 3
HELIX 46 48 3
HELIX 51 63 13
TURN 64 65 2
HELIX 69 71 3
STRAND 73 78 6
HELIX 80 82 3
TURN 85 87 3
HELIX 88 99 12
TURN 100 100 1
STRAND 104 109 6
STRAND 115 115 1
STRAND 124 124 1
TURN 126 127 2
STRAND 130 130 1
STRAND 132 132 1
HELIX 137 149 13
TURN 150 151 2
STRAND 153 159 7
HELIX 163 170 8
TURN 171 171 1
TURN 173 174 2
Internet
Start
Microsoft Word - Bioinfo1
NiceProt View of Swi...
11:33 μμ

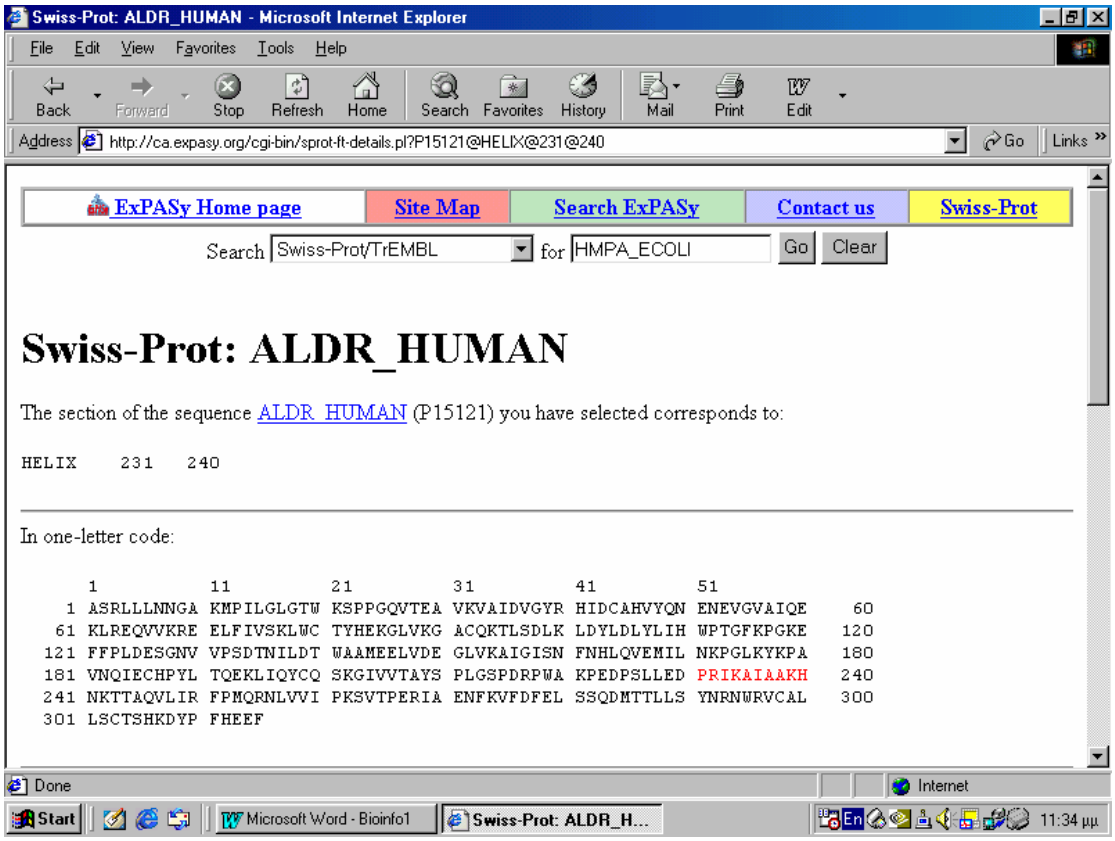

NiceProt View of Swiss-Prot: P15121 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/niceprot.pl?ALDR_HUMAN
Go Links
HELIX 163 170 8
TURN 171 171 1
TURN 173 174 2
STRAND 181 185 5
STRAND 187 187 1
TURN 188 189 2
STRAND 190 190 1
HELIX 193 201 9
TURN 202 203 2
STRAND 205 209 5
TURN 211 212 2
TURN 215 216 2
TURN 218 219 2
TURN 228 229 2
HELIX 231 240 10
TURN 241 241 1
HELIX 244 253 10
TURN 254 256 3
STRAND 258 259 2
STRAND 262 262 1
HELIX 266 273 8
HELIX 282 290 9
Internet
Start
Microsoft Word - Bioinfo1
NiceProt View of Swi...
11:33 μμ

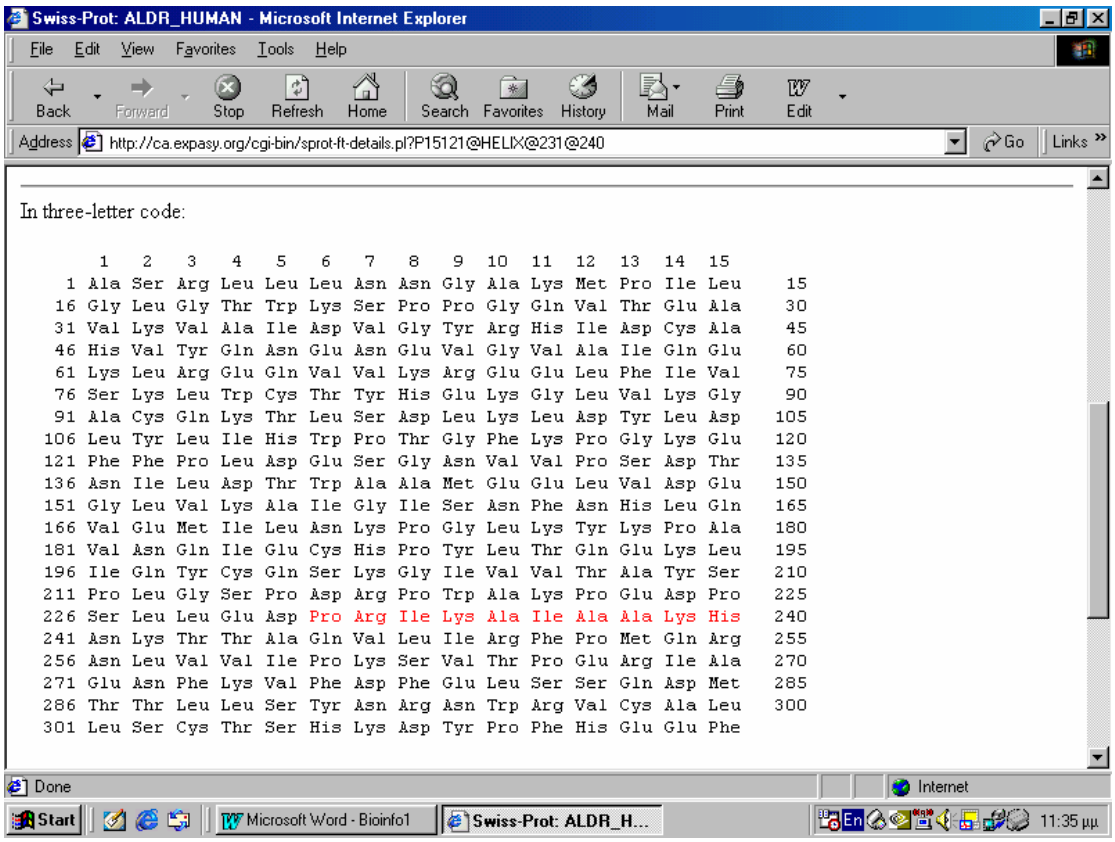
Κοιτώντας στην ενότητα των χαρακτηριστικών βλέπουμε ότι, για παράδειγμα, στη θέση 231-240 aa της ALDR_HUMAN υπάρχει ένας ΕΛΙΚΑΣ (HELIX). Πατώντας σε αυτό το τμήμα ανοίγει, μέσω της SWISS-PROT, η αλληλουχία της πρωτεΐνης σε κώδικα ενός γράμματος (one-letter code) ή τριών γραμμάτων (three-letter code), και με τον έλικα τονισμένο.






NiceProt View of Swiss-Prot: P15121 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/niceprot.pl?ALDR_HUMAN
Go Links
Sequence information
Length: 315 AA Molecular weight: 35722 Da CRC64: 99ABD1838964DCAE [This is a checksum on the sequence]
10 20 30 40 50 60
ASRLLLNNGA KMPILGLGTW KSPPGQVTEA VKVAIDVGYR HIDCAHVYQN ENEVGVAIQE
70 80 90 100 110 120
KLREQVVKRE ELFIVSKLWC TYHEKGLVKG ACQKTLSDLK LDYLDLYLIH WPTGFKPGKE
130 140 150 160 170 180
FFPLDESGNV VPSDTNILDT WAAMEELVDE GLVKAIGISN FNHLQVEMIL NKPGLKYKPA
190 200 210 220 230 240
VNQIECHPYL TQEKLIQYCQ SKGIVVTAYS PLGSPDRPWA KPEDPSLLED PRIKAIAAKH
250 260 270 280 290 300
NKTTAQVLIR FPMQRNLVVI PKSVTPERIA ENFKVFDFEL SSQDMTTLLS YNRNWRVCAL
310
LSCTSHKDYP FHEEF
Internet
Start
Microsoft Word - Bioinfo1
NiceProt View of Swi...
11:33 μμ


Swiss-Prot: ALDR_HUMAN - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/sprot-ft-details.pl?P15121@HELIX@231@240
Go Links
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Search Swiss-Prot/TrEMBL for HMPA_ECOLI Go Clear
Swiss-Prot: ALDR_HUMAN
The section of the sequence ALDR_HUMAN (P15121) you have selected corresponds to:
HELIX 231 240
In one-letter code:
1 11 21 31 41 51
1 ASRLLLNNGA KMPILGLGTW KSPPGQVTEA VKVAIDVGYR HIDCAHVYQN ENEVGVAIQE 60
61 KLREQVVKRE ELFIVSKLWC TYHEKGLVKG ACQKTLSDLK LDYLDLYLIH WPTGFKPGKE 120
121 FFPLDESGNV VPSDTNILDT WAAMEELVDE GLVKAIGISN FNHLQVEMIL NKPGLKYKPA 180
181 VNQIECHPYL TQEKLIQYCQ SKGIVVTAYS PLGSPDRPWA KPEDPSLLED PRIKAIAAKH 240
241 NKTTAQVLIR FPMQRNLVVI PKSVTPERIA ENFKVFDFEL SSQDMTTLLS YNRNWRVCAL 300
301 LSCTSHKDYP FHEEF
Done
Internet
Start
Microsoft Word - Bioinfo1
Swiss-Prot: ALDR_H...
11:34 μμ

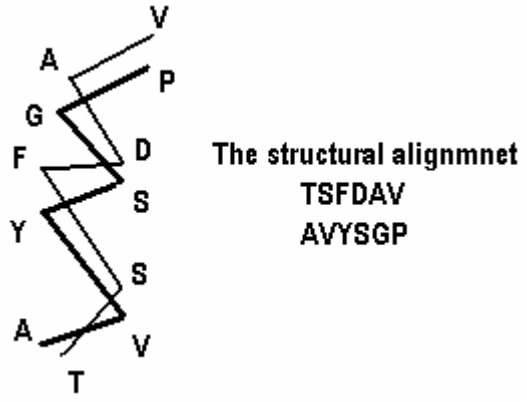

Swiss-Prot: ALDR_HUMAN - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://ca.expasy.org/cgi-bin/sprot-ft-details.pl?P15121@HELIX@231@240
Go Links
In three-letter code:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
1 Ala Ser Arg Leu Leu Leu Asn Asn Gly Ala Lys Met Pro Ile Leu 15
16 Gly Leu Gly Thr Trp Lys Ser Pro Pro Gly Gln Val Thr Glu Ala 30
31 Val Lys Val Ala Ile Asp Val Gly Tyr Arg His Ile Asp Cys Ala 45
46 His Val Tyr Gln Asn Glu Asn Glu Val Gly Val Ala Ile Gln Glu 60
61 Lys Leu Arg Glu Gln Val Val Lys Arg Glu Glu Leu Phe Ile Val 75
76 Ser Lys Leu Trp Cys Thr Tyr His Glu Lys Gly Leu Val Lys Gly 90
91 Ala Cys Gln Lys Thr Leu Ser Asp Leu Lys Leu Asp Tyr Leu Asp 105
106 Leu Tyr Leu Ile His Trp Pro Thr Gly Phe Lys Pro Gly Lys Glu 120
121 Phe Phe Pro Leu Asp Glu Ser Gly Asn Val Val Pro Ser Asp Thr 135
136 Asn Ile Leu Asp Thr Trp Ala Ala Met Glu Glu Leu Val Asp Glu 150
151 Gly Leu Val Lys Ala Ile Gly Ile Ser Asn Phe Asn His Leu Gln 165
166 Val Glu Met Ile Leu Asn Lys Pro Gly Leu Lys Tyr Lys Pro Ala 180
181 Val Asn Gln Ile Glu Cys His Pro Tyr Leu Thr Gln Glu Lys Leu 195
196 Ile Gln Tyr Cys Gln Ser Lys Gly Ile Val Val Thr Ala Tyr Ser 210
211 Pro Leu Gly Ser Pro Asp Arg Pro Trp Ala Lys Pro Glu Asp Pro 225
226 Ser Leu Leu Glu Asp Pro Arg Ile Lys Ala Ile Ala Ala Lys His 240
241 Asn Lys Thr Thr Ala Gln Val Leu Ile Arg Phe Pro Met Gln Arg 255
256 Asn Leu Val Val Ile Pro Lys Ser Val Thr Pro Glu Arg Ile Ala 270
271 Glu Asn Phe Lys Val Phe Asp Phe Glu Leu Ser Ser Gln Asp Met 285
286 Thr Thr Leu Leu Ser Tyr Asn Arg Asn Trp Arg Val Cys Ala Leu 300
301 Leu Ser Cys Thr Ser His Lys Asp Tyr Pro Phe His Glu Glu Phe
Done
Internet
Start
Microsoft Word - Bioinfo1
Swiss-Prot: ALDR_H...
11:35 μμ

Σύγκριση δομών πρωτεϊνών

Η σύγκριση δομών πρωτεϊνών είναι χρήσιμη για την απόδειξη μακρινών εξελικτικών σχέσεων. Οι σχέσεις αυτές δεν είναι ανιχνεύσιμες με τις μεθόδους οι οποίες βασίζονται στις αλληλουχίες, καθώς η εξέλιξη έχει μικρότερη επίδραση και επομένως και αλλάζει σε μικρότερο βαθμό τη δομή απ' ό,τι την αλληλουχία των πρωτεϊνών. Μπορεί επίσης να βοηθήσει και στο σχεδιασμό φαρμάκων καθώς βοηθά στην ανάλυση διαμορφωτικών μεταβολών στη θέση με την οποία ο συνθέτης συνδέεται με το υπόστρωμα (conformational changes on ligand binding).
Μέσα σε μία οικογένεια πρωτεϊνών, η σύγκριση των δομών τους δίνει πολύτιμα στοιχεία σχετικά με την ανοχή της δεδομένης οικογένειας σε δομικές αλλαγές, καθώς επίσης και τις επιπτώσεις που μπορεί να έχει η οποιαδήποτε αλλαγή στις λειτουργίες των πρωτεϊνών.

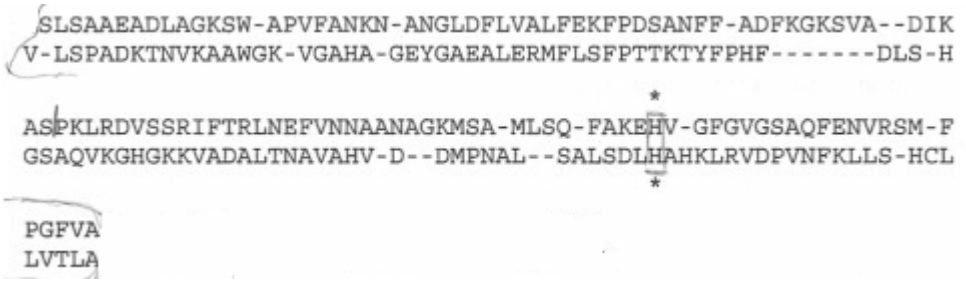
Αντιστοίχηση δομών

Σε γενικές γραμμές, η αντιστοίχηση δομών είναι μία διαδικασία κατά την οποία δύο παρόμοιες, τρισδιάστατες δομές τοποθετούνται η μία πάνω στην άλλη, έτσι ώστε οι πεπτιδική αλυσίδα των δομικώς όμοιων αμινοξέα να συμπίπτουν όσο το δυνατόν περισσότερο στο χώρο. Αυτή η σύγκριση στη συνέχεια χρησιμοποιείται για τον καθορισμό μίας αντιστοίχησης αλληλουχιών, στην οποία τα αντιστοιχιζόμενα αμινοξέα μοιάζουν δομικά.



Λόγω του ότι κατά την πορεία της εξέλιξης η δομή διατηρείται καλύτερα απ' ό,τι η αλληλουχία, η αντιστοίχηση δομών έχει περισσότερες πιθανότητες να είναι σωστή όσον αφορά τη βιολογική λειτουργία και την εξέλιξη.

Ακολούθως παρουσιάζονται δύο αντιστοιχήσεις ενός ζεύγους αλληλουχιών σφαιρίνης (globin) με μακρινή σχέση: (a) η πρώτη αντιστοίχηση προέκυψε χρησιμοποιώντας τις αλληλουχίες και (b) η δεύτερη είναι αντιστοίχηση δομών. Οι δύο αντιστοιχήσεις, όπως είναι αναμενόμενο, διαφέρουν, με την αντιστοίχηση αλληλουχιών να αποδίδει μεγαλύτερη ομοιότητα ανάμεσά τους. Ωστόσο, τα histine (H) αμινοξέα αντιστοιχίστηκαν (εξισώθηκαν) μόνο στην αντιστοίχηση δομών. Η histine είναι σημαντική, καθώς ελέγχει το σίδηρο (heme iron) και στις δύο σφαιρίνες.



Πηγές δεδομένων (Data resources)– PDB
Ο μεγαλύτερος αριθμός δομών πρωτεϊνών βρίσκεται αποθηκευμένος στην Protein Databank (PDB), η οποία συντηρείται στο Research Collaboratory of Structural Biology (RCSB), του Rutgers University και στο EBI.

Αλγόριθμοι
Η πλειονότητα των χρησιμοποιούμενων μεθόδων βασίζεται στη σύγκριση των γεωμετρικών ιδιοτήτων (properties) και σχέσεων ανάμεσα στα στοιχεία της δευτεροταγούς δομής και των αμινοξέα που βρίσκονται κατά μήκος της ανθρακικής αλυσίδας (χρησιμοποιούνται άτομα Ca και Cb). Οι γεωμετρικές ιδιότητες των αμινοξέα και οι δευτεροταγείς δομές καθορίζονται από τις τρισδιάστατες συντεταγμένες της δομής, οι οποίες βρίσκονται καταχωρημένες στην PDB. Στις πληροφορίες για τις σχέσεις τους περιλαμβάνονται οι αποστάσεις των διανυσμάτων ανάμεσα σε θέσεις της δευτεροταγούς δομής ή ανάμεσα σε αμινοξέα. Ορισμένες φορές στη σύγκριση περιλαμβάνονται και άλλες ιδιότητες, όπως για παράδειγμα φυσικοχημικές ιδιότητες (π.χ. υδροφοβία).

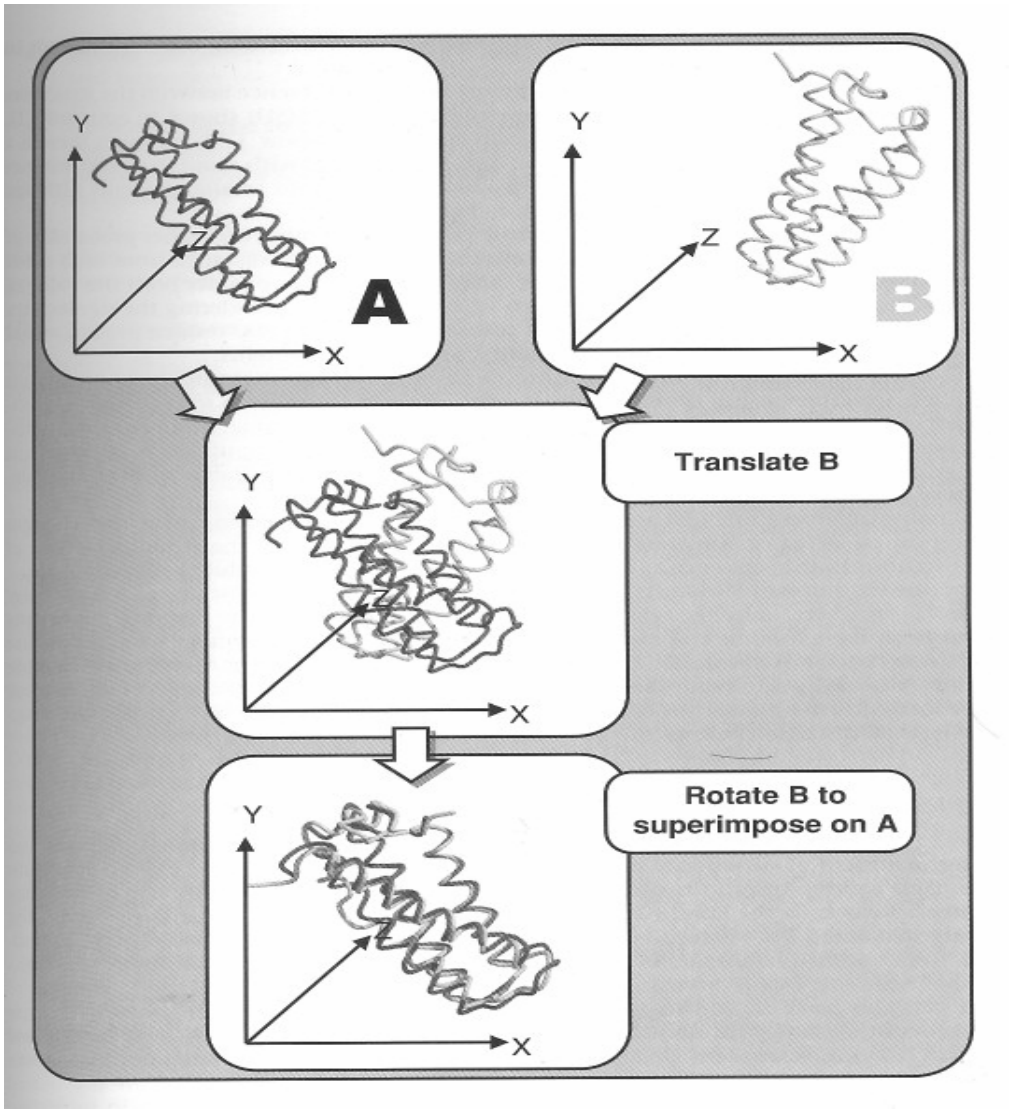
Μέθοδοι
Οι μέθοδοι για τη σύγκριση δευτεροταγών δομών χωρίζονται σε δύο κατηγορίες: Ενδομοριακές (intermolecular) και διαμοριακές (intramolecular) μέθοδοι.

Ενδομοριακές (intermolecular) μέθοδοι
Οι ενδομοριακές μέθοδοι (rigid body superposition methods) συγκρίνουν, τοποθετώντας τη μία πάνω στην άλλη, δομές πρωτεϊνών και υπολογίζουν τις ενδομοριακές αποστάσεις. Οι μέθοδοι αυτές συγκρίνουν γεωμετρικές ιδιότητες, π.χ. τις θέσεις των αμινοξέα σε τρισδιάστατο χώρο.

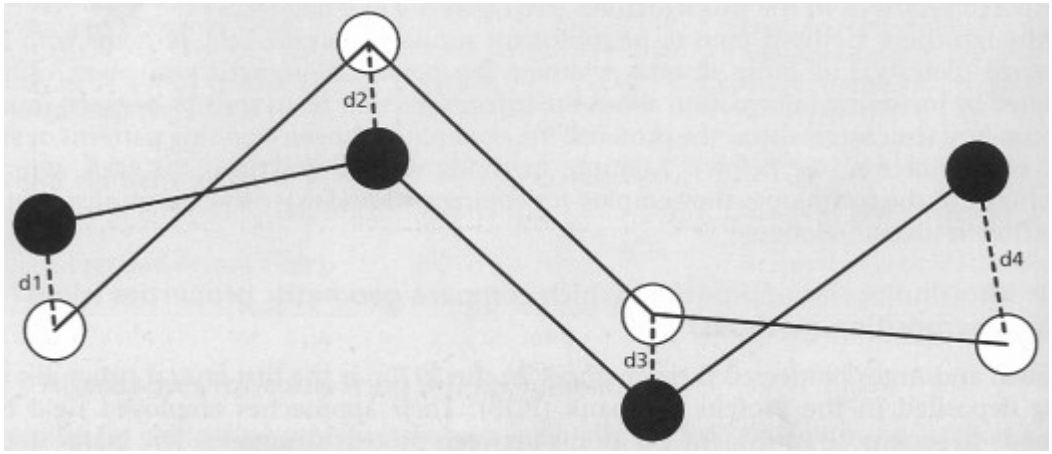
Τα βασικά βήματα της μεθόδου αυτής είναι:

1- Μετάφραση και των δύο πρωτεϊνών σε μία κοινή θέση πάνω στο τριών διαστάσεων πλαίσιο αναφοράς.
2- Περιστροφή της μίας πρωτεΐνης με σημείο αναφοράς τη δεύτερη πρωτεΐνη, γύρω από τους τρεις βασικούς άξονες.
3- Υπολογισμός των αποστάσεων ανάμεσα σε αντίστοιχες θέσεις (equivalent positions) των δύο πρωτεϊνών.

Επανάληψη των βημάτων 2 και 3 έως ότου επιτευχθεί σύγκλιση στη μικρότερη δυνατή απόσταση ανάμεσα στις δύο συγκρινόμενες δομές.



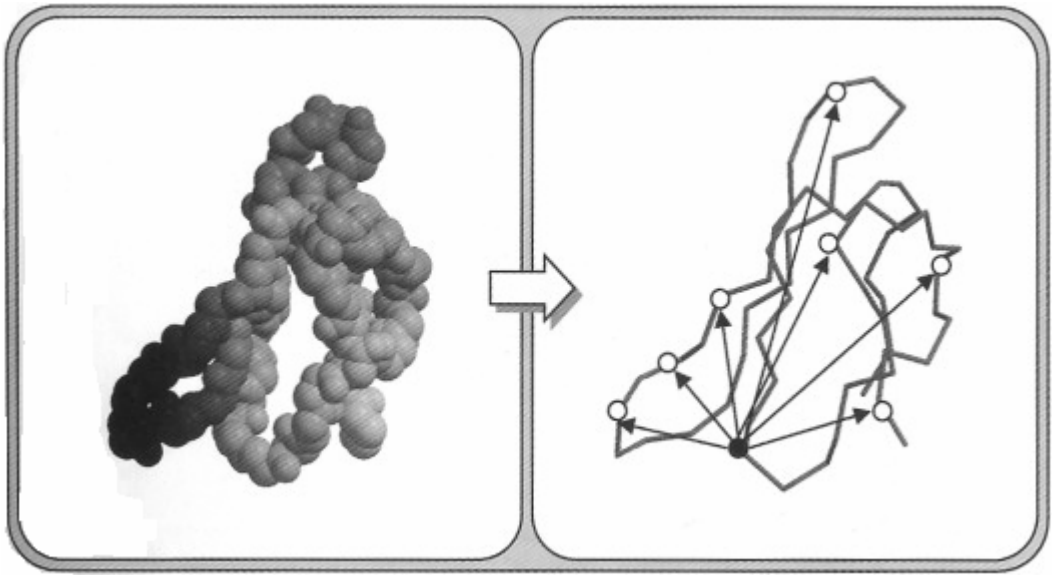
Η απόσταση ανάμεσα σε δύο αντίστοιχες θέσεις περιγράφεται με τη βοήθεια μίας συνάρτηση διαφορών (residual function) , η οποία υπολογίζει την απόσταση ανάμεσα σε συγκρινόμενα αμινοξέα και η οποία είναι η εξίσωση των ελαχίστων τετραγώνων (Root Mean Squared Deviation – RMSD): $RMSD=\sqrt{\Sigma d_i^2}/N$.

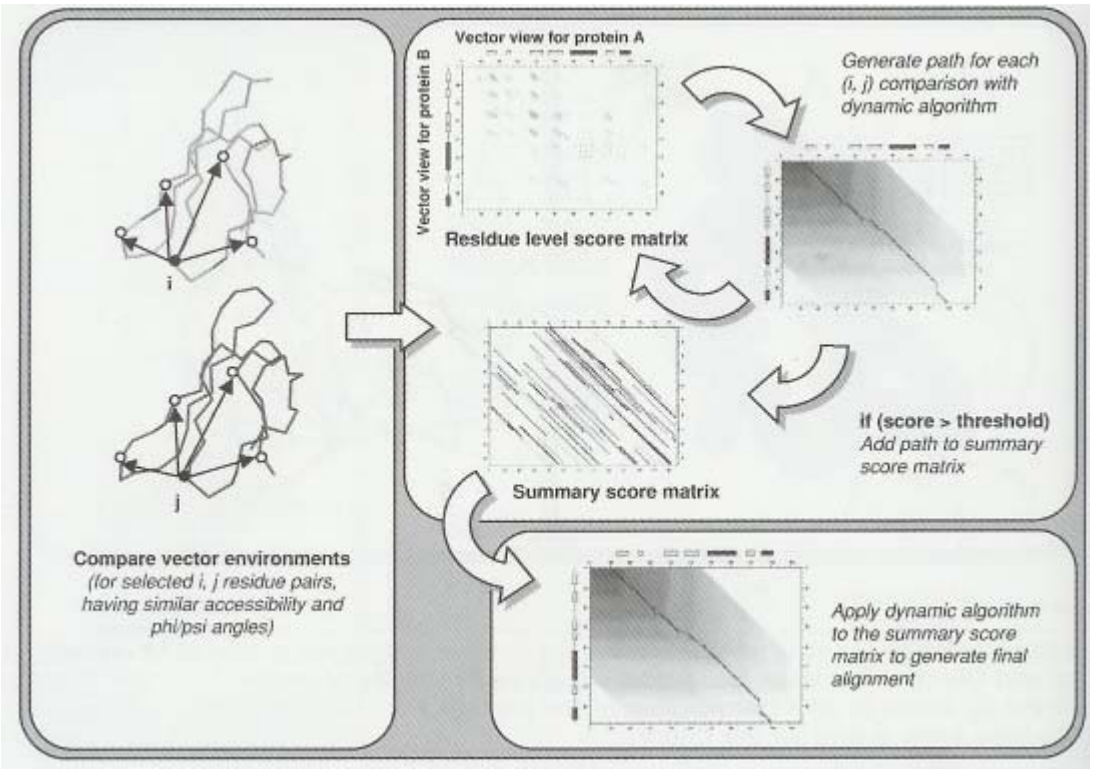


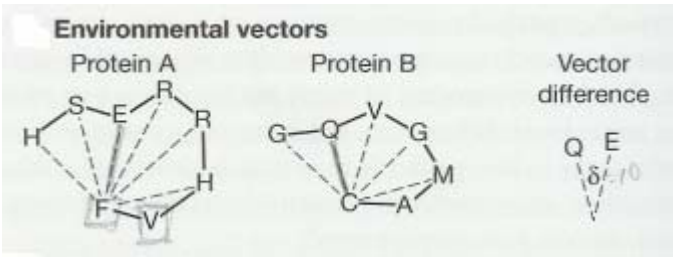
Διαμοριακές μέθοδοι

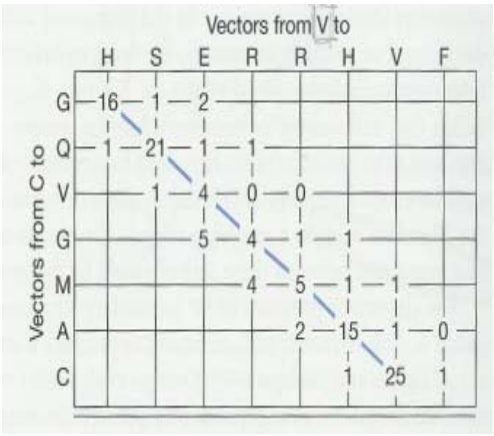
Οι διαμοριακές μέθοδοι βασίζονται στη σύγκριση διαμοριακών αποστάσεων ή διανυσμάτων. Αντιστοιχίζουν δομές πρωτεϊνών με βάση πληροφορίες σχετικά με εσωτερικές σχέσεις μέσα σε κάθε πρωτεΐνη, π.χ. αποστάσεις μεταξύ αμινοξέα ή δευτεροταγείς δομές μέσα σε μία πρωτεΐνη.

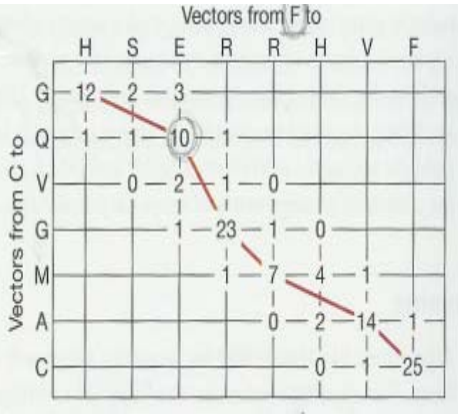
Η μέθοδος SSAP είναι μια διαμοριακή μέθοδος για τη σύγκριση δύο δομών πρωτεϊνών, η οποία εφαρμόζει δυναμικό προγραμματισμό σε δύο επίπεδα.

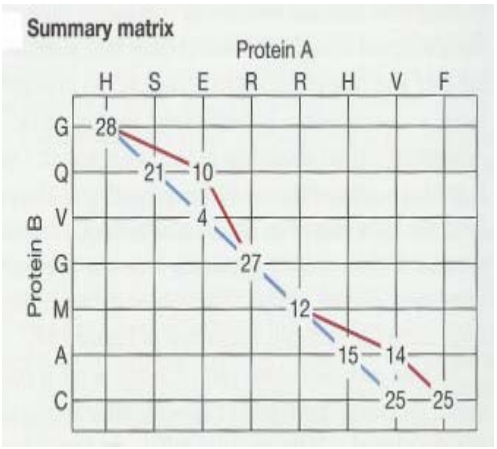












Περιληπτικά, το SSAP περιλαμβάνει τα ακόλουθα βήματα:

1- Υπολογισμό της οπτικής των διανυσμάτων (vector view) για κάθε residue στις δύο πρωτεΐνες (π.χ. για το residue i για τη μία πρωτεΐνη και για το residue j για την άλλη). Το vector view (ή δομικό περιβάλλον του residue – residue structural environment) ορίζεται ως το σύνολο των διανυσμάτων από το residue i προς όλα τα άλλα αμινοξέα στην πρωτεΐνη.

2- Σύγκριση ζευγών αμινοξέα ανάμεσα στις πρωτεΐνες προκειμένου να βρεθούν ενδεχόμενες ομοιότητες ανάμεσα στα αμινοξέα (π.χ. η ύπαρξη παρόμοιων γωνιών στρέψης και περιοχών πρόσβασης). Η βαθμολόγηση των συγκρίσεων ανάμεσα σε δύο vector views γίνεται σύμφωνα με την ομοιότητα των διανυσμάτων, σε ένα δισδιάστατο πίνακα, ο οποίος ονομάζεται πίνακας βαθμολογίας του επιπέδου των αμινοξέα (residue level score matrix). Για την εύρεση του βέλτιστου μονοπατιού μέσα τον πίνακα χρησιμοποιείται δυναμικός προγραμματισμός, δίνοντας τη βέλτιστη αντιστοίχηση των residue views.

3- Για περιπτώσεις ζευγών αμινοξέα τα οποία παίρνουν υψηλή βαθμολογία, τα μονοπάτια που προκύπτουν κατά το βήμα 2 συγκεντρώνονται /

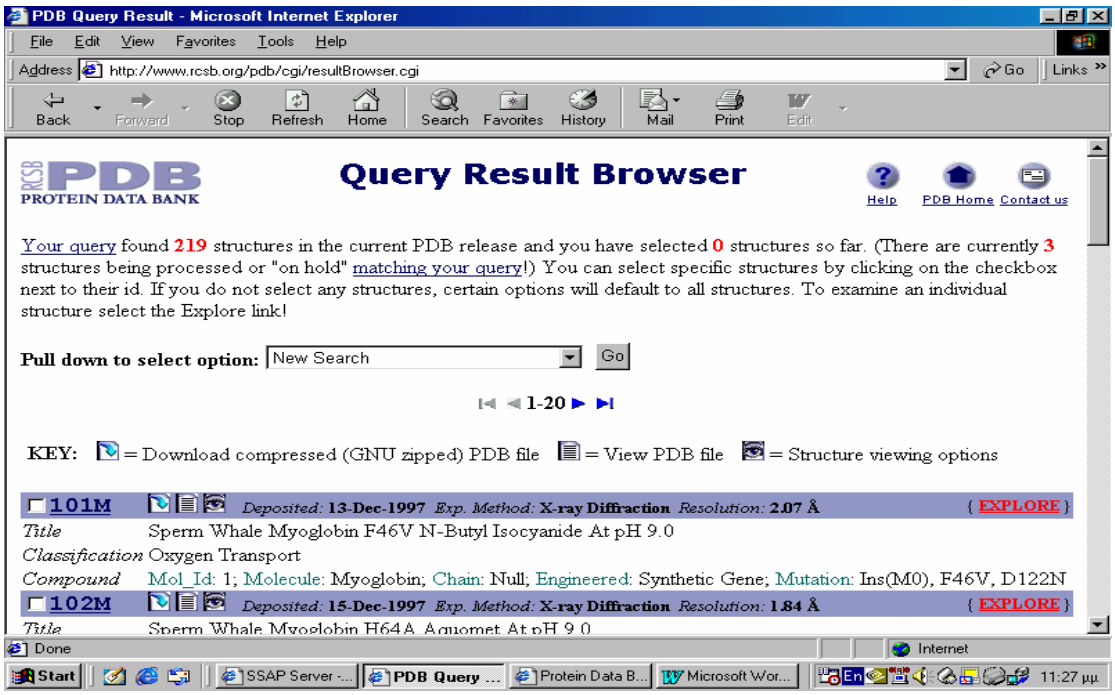
συγχωνεύονται σε ένα δισδιάστατο συγκεντρωτικό πίνακα βαθμολογίας (2D summary score matrix).

4- Επανάληψη των βημάτων 2 και 3 έως ότου όλα τα ενδεχομένως ισοδύναμα ζευγάρια να έχουν συγκριθεί.
5- Και πάλι χρησιμοποίηση δυναμικού προγραμματισμού προκειμένου να καθοριστεί το βέλτιστο μονοπάτι μέσα στο δισδιάστατο συγκεντρωτικό πίνακα βαθμολογίας.

ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

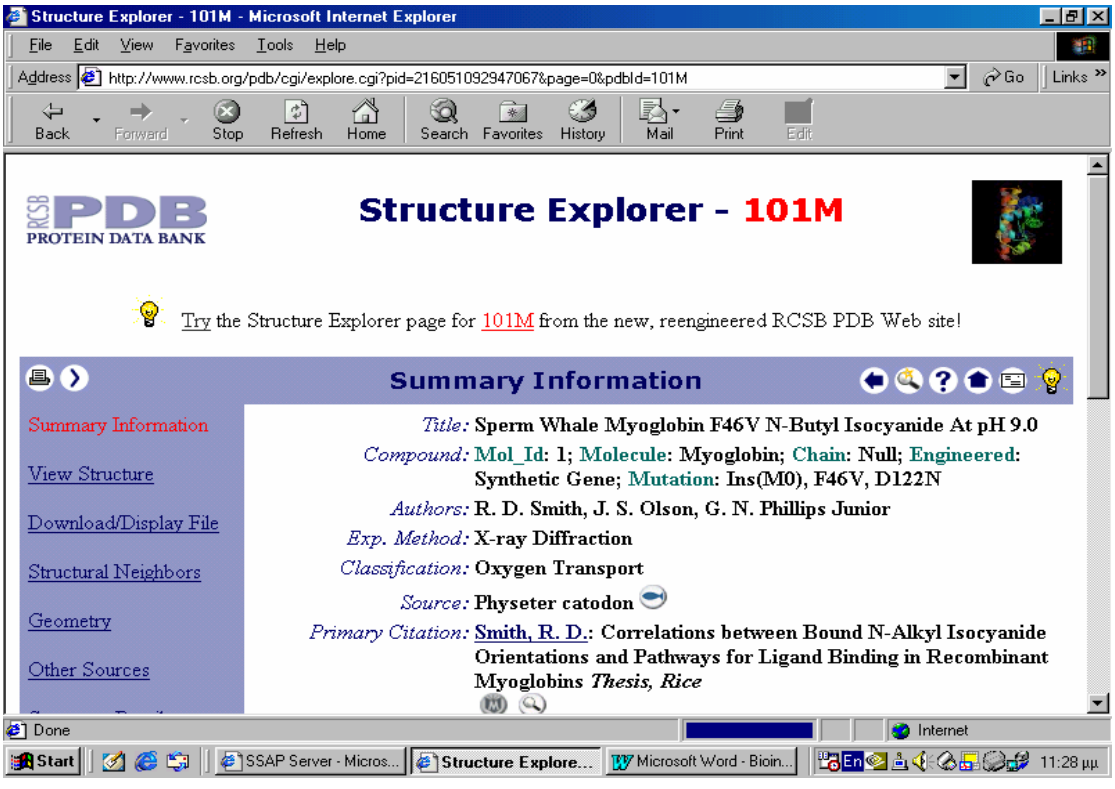
Σύγκριση δομών πρωτεϊνών – Μέθοδος SSAP

Προκειμένου να συγκρίνουμε τις δομές δύο πρωτεϊνών, π.χ. δύο myoglobins, αρχικά πρέπει να προσδιορίσουμε τους κωδικούς που τους αντιστοιχούν στην PDB.
Για να επισκεφθείτε την ιστοσελίδα της PDB, ανοίξτε τον Internet Explorer και πληκτρολογήστε τη διεύθυνση www.rcsb.org/pdb/ . Στη συνέχεια πληκτρολογήστε στο κενό πεδίο τη λέξη 'myoglobin' και πατήστε 'Search'.
Αυτό θα σας οδηγήσει σε μία νέα σελίδα όπου παραθέτονται όλες οι myoglobins

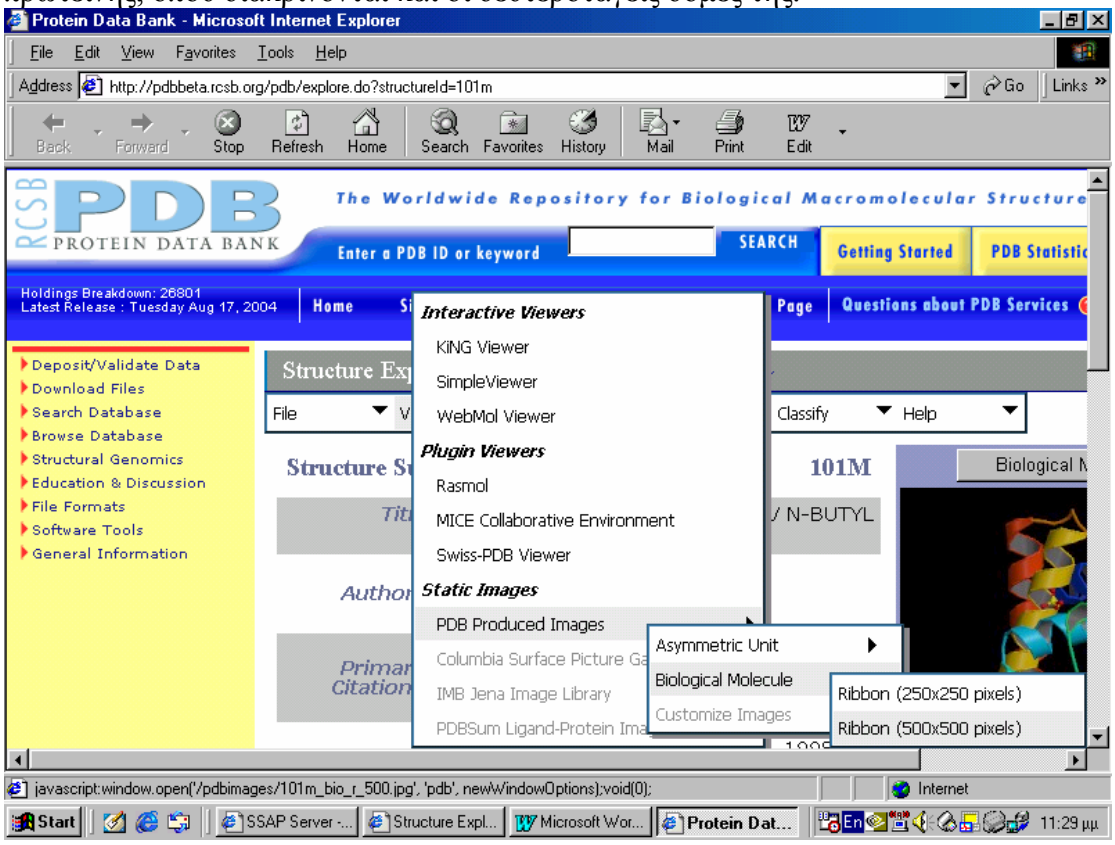


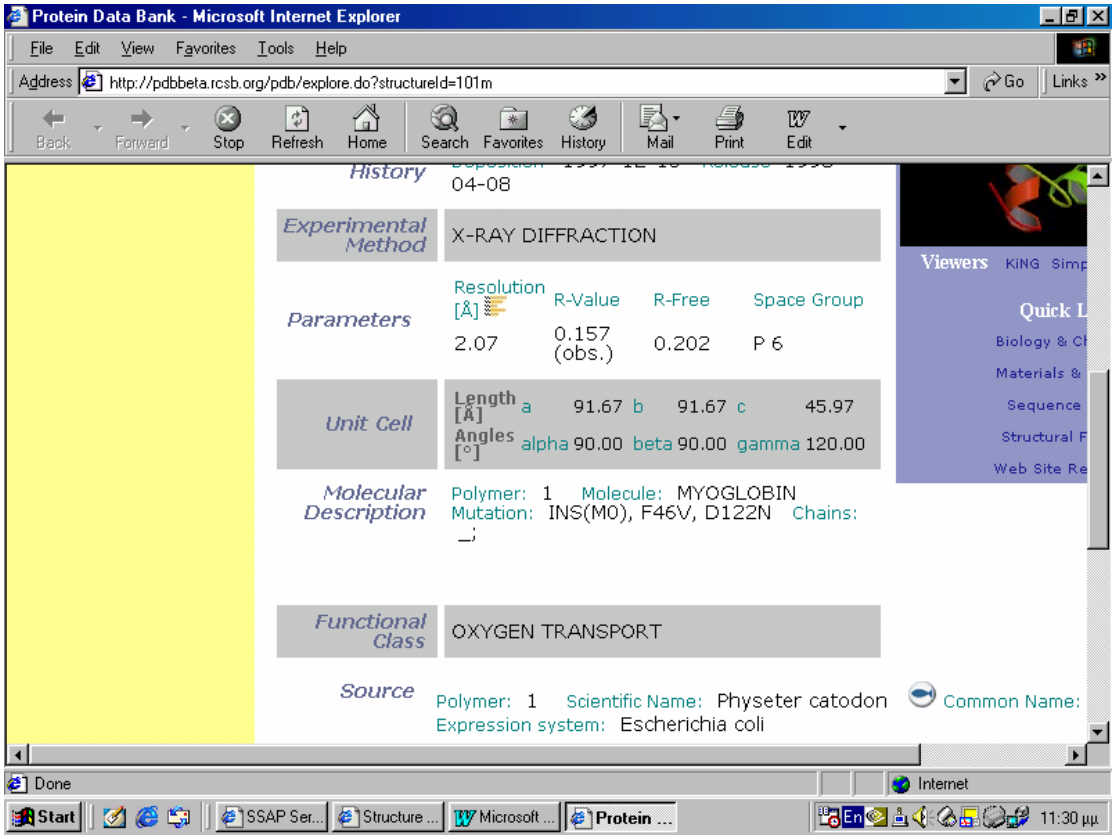
συνοδευόμενες από τους κωδικούς τους και όλες τις πληροφορίες σχετικά με τις πρωτεΐνες.

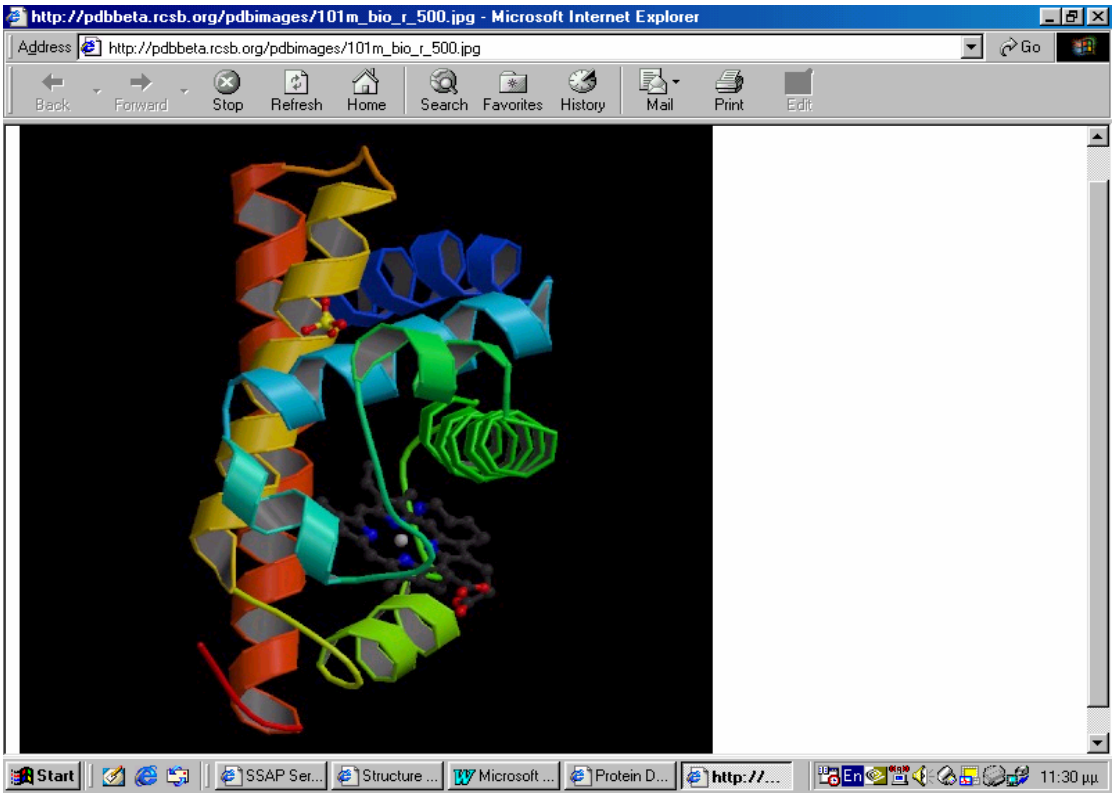
Πατώντας στον κωδικό πρωτεΐνης '101M', εμφανίζονται περιληπτικές πληροφορίες σχετικά με αυτήν την πρωτεΐνη.



Πατώντας εν συνεχεία στο link ‘View Structure’, που βρίσκεται στα αριστερά της σελίδας, και την επιλογή ‘Ribbon’ μπορούμε να δούμε μία τρισδιάστατη εικόνα της πρωτεΐνης, όπου διακρίνονται και οι δευτεροταγείς δομές της.

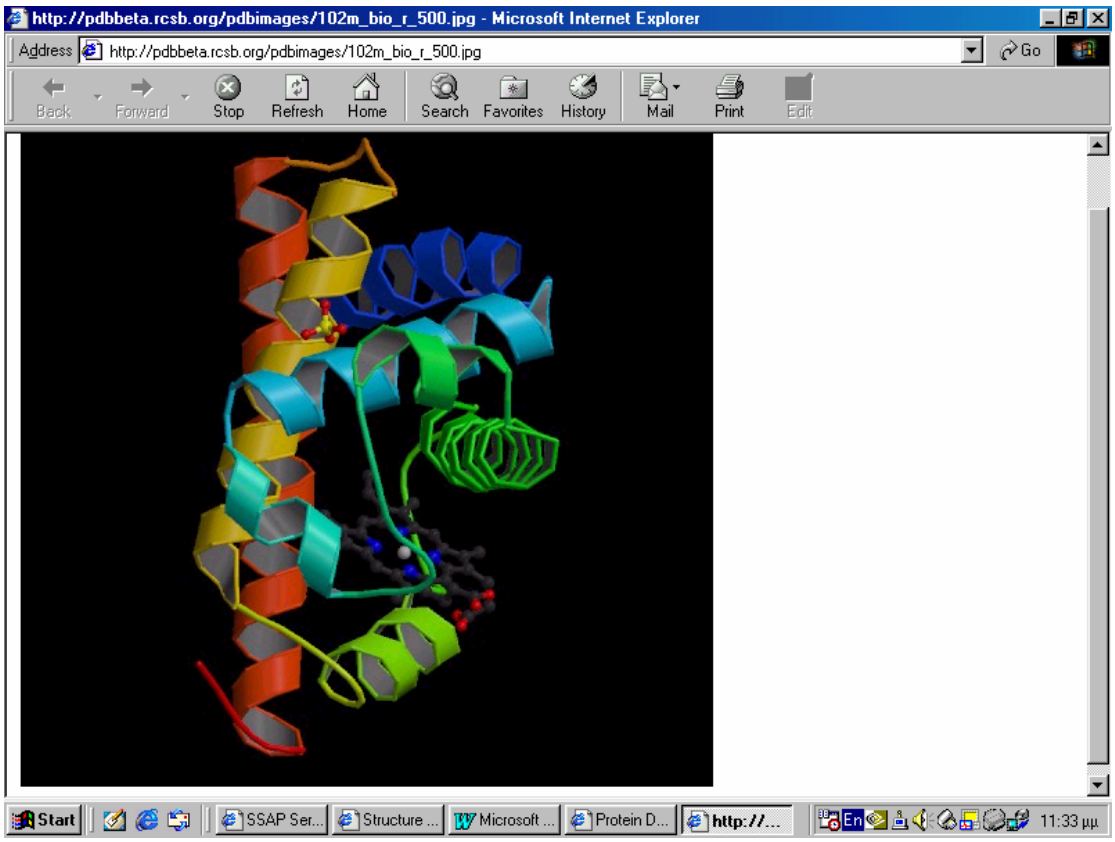




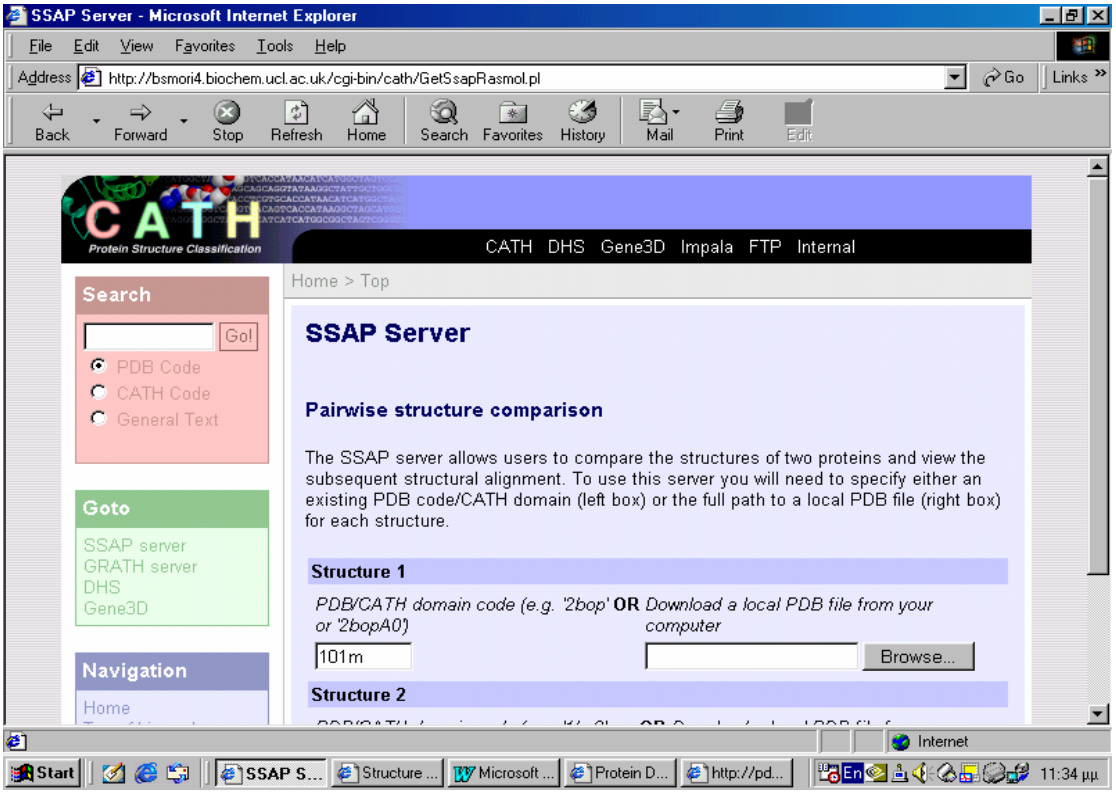


Πατώντας στον κωδικό πρωτεΐνης ‘102M’, και μετά στο ‘View Structure’ μπορούμε να δούμε πληροφορίες για μία άλλη myoglobin. Αν συγκρίνουμε τις εικόνες για τις

δύο αυτές myoglobins (την 101M και την 102M) παρατηρούμε ότι παρουσιάζουν πολλές ομοιότητες ως προς τη δομή τους.



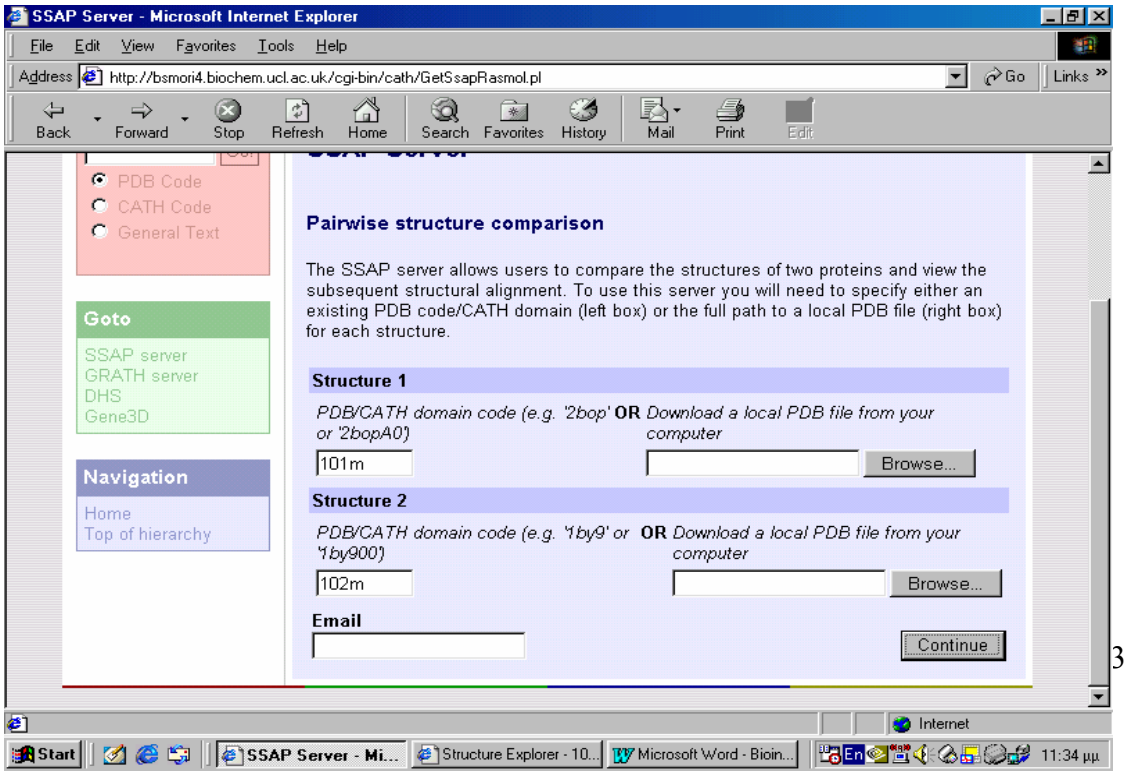
Για να συγκρίνουμε τώρα τις δομές των δύο πρωτεϊνών 101M και 102Μα χρησιμοποιώντας τη μέθοδο SSAP, πληκτρολογήστε στον Internet Explorer τη

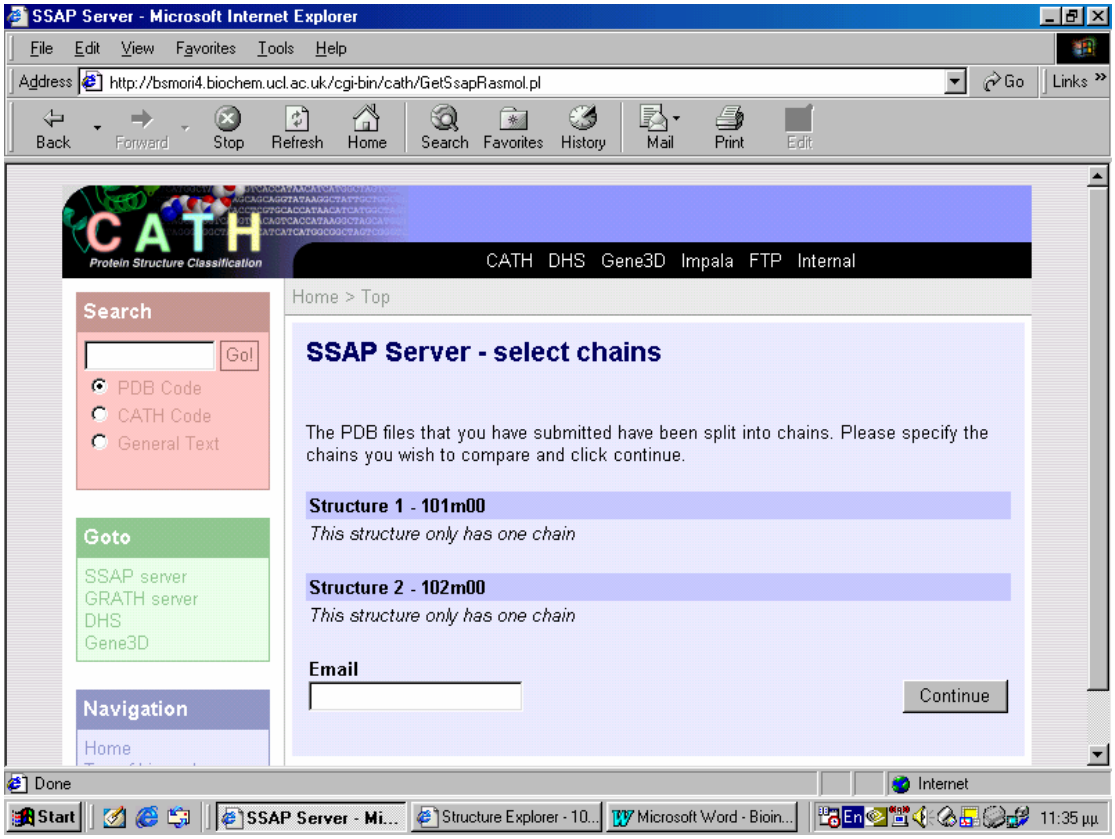


διεύθυνση: www.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl.
Στο κενό πεδίο κάτω από τον τίτλο 'Structure 1' πληκτρολογήστε 101m και στο πεδίο κάτω από το 'Structure 2' πληκτρολογήστε 102m. Έπειτα πατήστε 'Continue'..

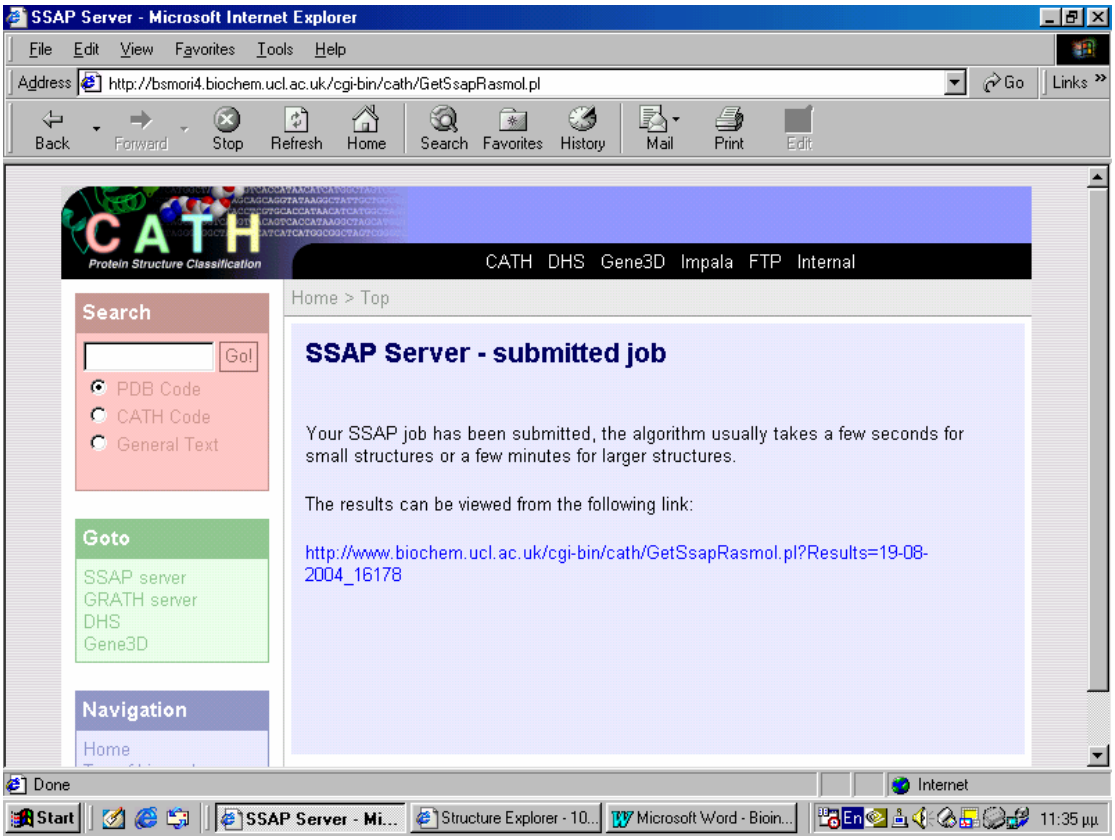
Παρουσιάζεται η δυνατότητα επιλογής για τις πρωτεϊνικές αλυσίδες προς σύγκριση, ωστόσο στη συγκεκριμένη περίπτωση του παραδείγματός μας, οι myoglobins έχουν μόνο μία αλυσίδα.

Για να δείτε τα αποτελέσματα πατήστε τη διεύθυνση που εμφανίζεται στο επόμενο

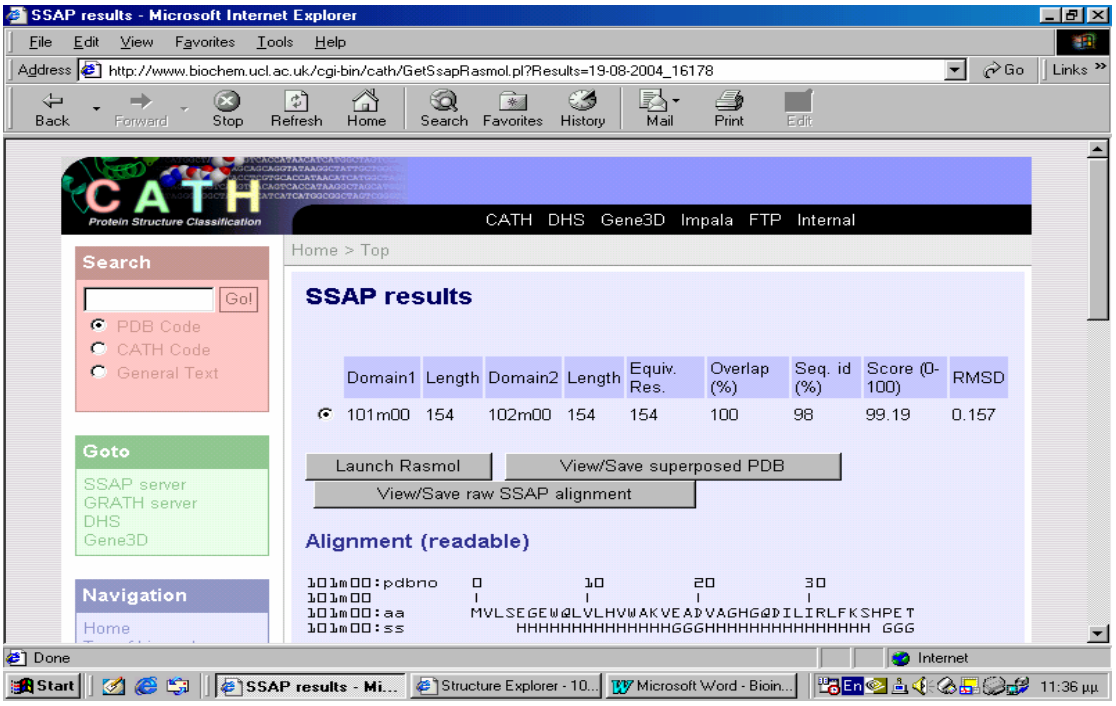


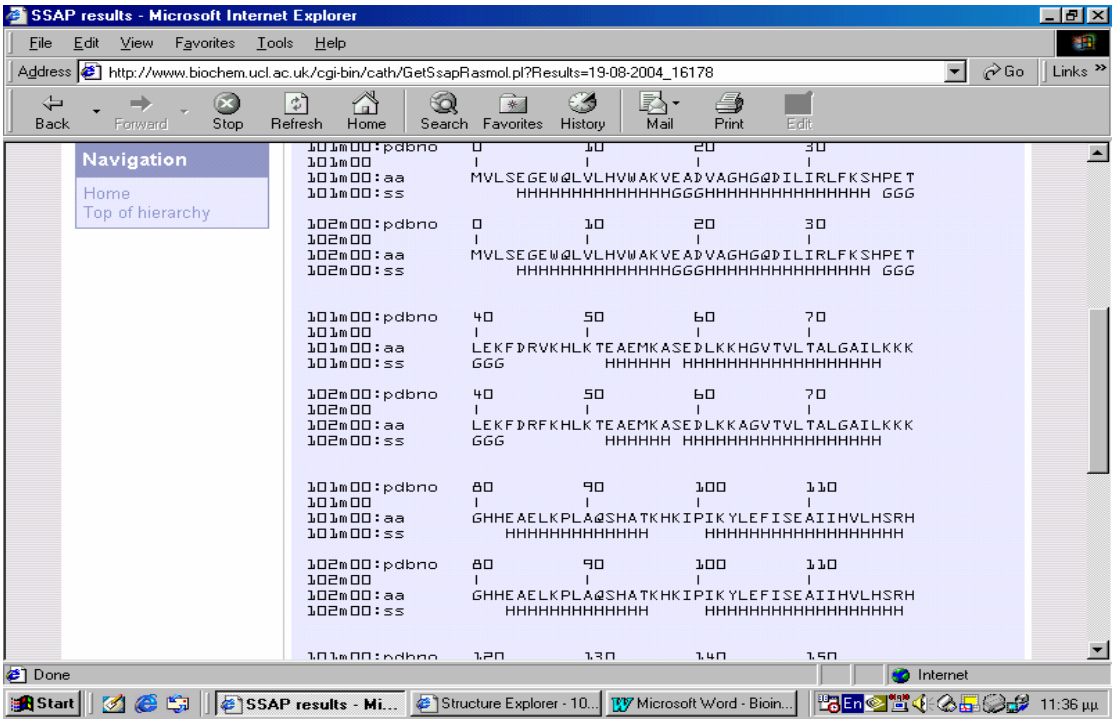

SSAP Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://bsmori4.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
CATH
Protein Structure Classification
CATH DHS Gene3D Impala FTP Internal
Search
Go!
PDB Code
CATH Code
General Text
Goto
SSAP server
GRATH server
DHS
Gene3D
Navigation
Home
Home > Top
SSAP Server - select chains
The PDB files that you have submitted have been split into chains. Please specify the chains you wish to compare and click continue.
Structure 1 - 101m00
This structure only has one chain
Structure 2 - 102m00
This structure only has one chain
Email
Continue
Done
Internet
Start
SSAP Server - Mi...
Structure Explorer - 10...
Microsoft Word - Bioin...
11:35 μμ

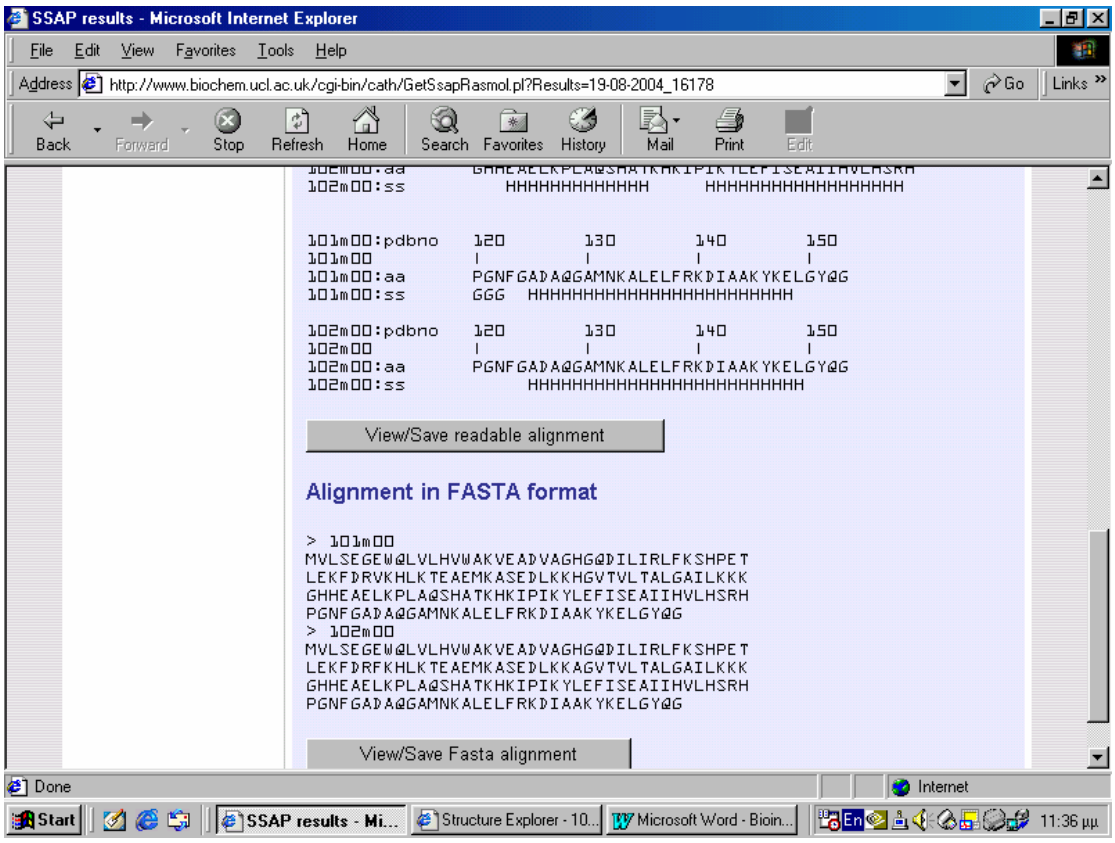
παράθυρο.


SSAP Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://bsmori4.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
CATH
Protein Structure Classification
CATH DHS Gene3D Impala FTP Internal
Search
Go!
PDB Code
CATH Code
General Text
Goto
SSAP server
GRATH server
DHS
Gene3D
Navigation
Home
Home > Top
SSAP Server - submitted job
Your SSAP job has been submitted, the algorithm usually takes a few seconds for small structures or a few minutes for larger structures.
The results can be viewed from the following link:
http://www.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl?Results=19-08-2004_16178
Done
Internet
Start
SSAP Server - Mi...
Structure Explorer - 10...
Microsoft Word - Bioin...
11:35 μμ

Εκεί παρατίθεται με το RMSD η αντιστοίχηση των δομών για τις δύο πρωτεΐνες, η βαθμολογία ταύτισης και η ομοιότητα των αλληλουχιών. Επιπλέον, παρουσιάζεται ενδεικτικά και η δευτεροταγής δομή (secondary structure – ss).



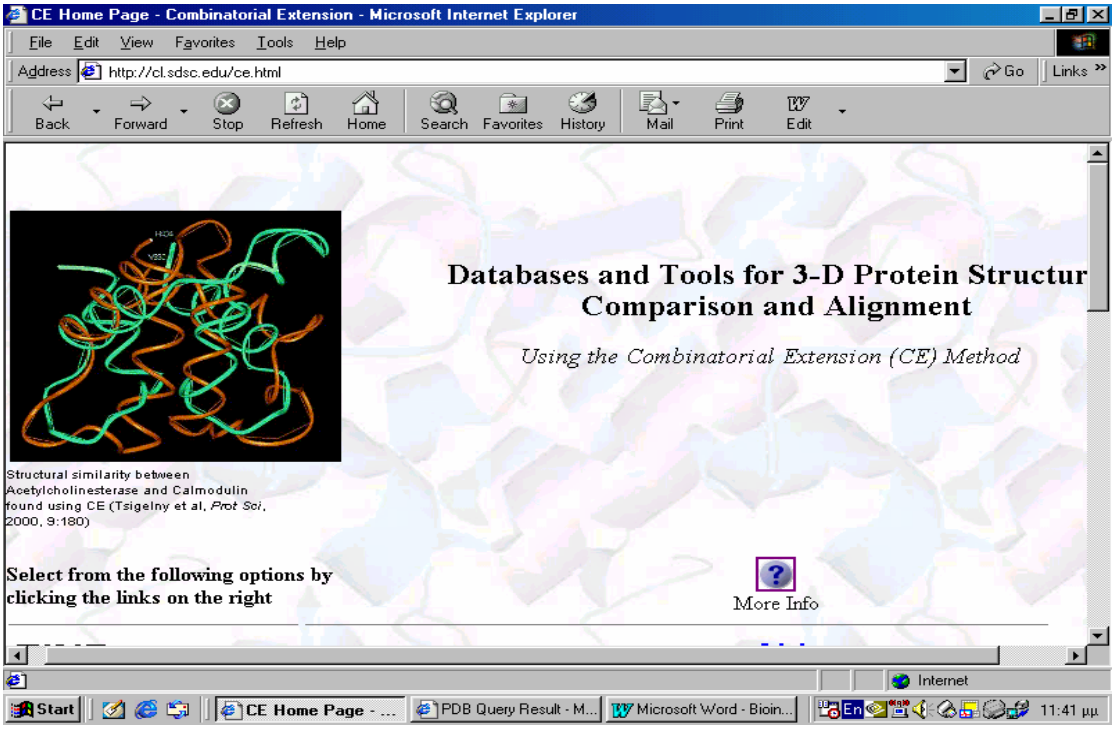



SSAP results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl?Results=19-08-2004_16178
Go Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
1o2mOO:ss HHHHHHHHHHHHH HHHHHHHHHHHHHHHHHH
1o1mOO:pdbno 120 130 140 150
1o1mOO
1o1mOO:aa PGNFGADA@GAMNKALELFRKDIAAKYKELGY@G
1o1mOO:ss GGG HHHHHHHHHHHHHHHHHHHHHHHH
1o2mOO:pdbno 120 130 140 150
1o2mOO
1o2mOO:aa PGNFGADA@GAMNKALELFRKDIAAKYKELGY@G
1o2mOO:ss HHHHHHHHHHHHHHHHHHHHHHHHH
View/Save readable alignment
Alignment in FASTA format
> 1o1mOO
MVLSEGEW@LVLHVWAKVEADVAGHG@DILIRLFKSHPET
LEKFDRVKHLKTEAEMKASEDLKKHGVTVLTALGAILKKK
GHHEAELKPLA@SHATKHKIPIKYLEFISEAIIHVLHSRH
PGNFGADA@GAMNKALELFRKDIAAKYKELGY@G
> 1o2mOO
MVLSEGEW@LVLHVWAKVEADVAGHG@DILIRLFKSHPET
LEKFDRFKHLKTEAEMKASEDLKKAGVTVLTALGAILKKK
GHHEAELKPLA@SHATKHKIPIKYLEFISEAIIHVLHSRH
PGNFGADA@GAMNKALELFRKDIAAKYKELGY@G
View/Save Fasta alignment
Done
Internet
Start
SSAP results - Mi...
Structure Explorer - 10...
Microsoft Word - Bioin...
11:36 μμ

Συνδυασμός ενδομοριακών και διαμοριακών μεθόδων σύγκρισης

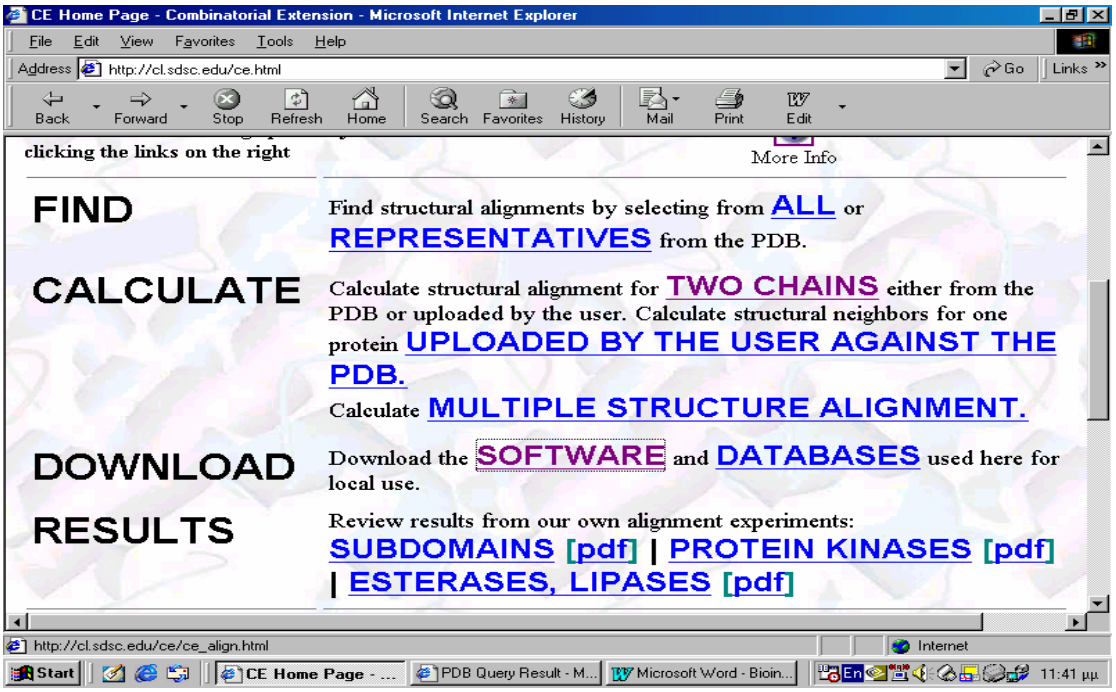
Μία άλλη μέθοδος για τη σύγκριση δομών πρωτεϊνών είναι η μέθοδος CE (Combinatorial Extension – Συνδυαστική Επέκταση), η οποία συνδυάζει τις ενδομοριακές και τις διαμοριακές μεθόδους (για λεπτομέρειες σχετικά με την μέθοδο CE συμβουλευτείτε την βοήθεια που ακολουθεί, ή το άρθρο του Tsigelny, Prot. Sci. 2000).

Για να επισκεφθείτε την ιστοσελίδα με πρόσβαση στη βάση δεδομένων για τη CE, ανοίξτε τον Internet Explorer και πληκτρολογήστε τη διεύθυνση:

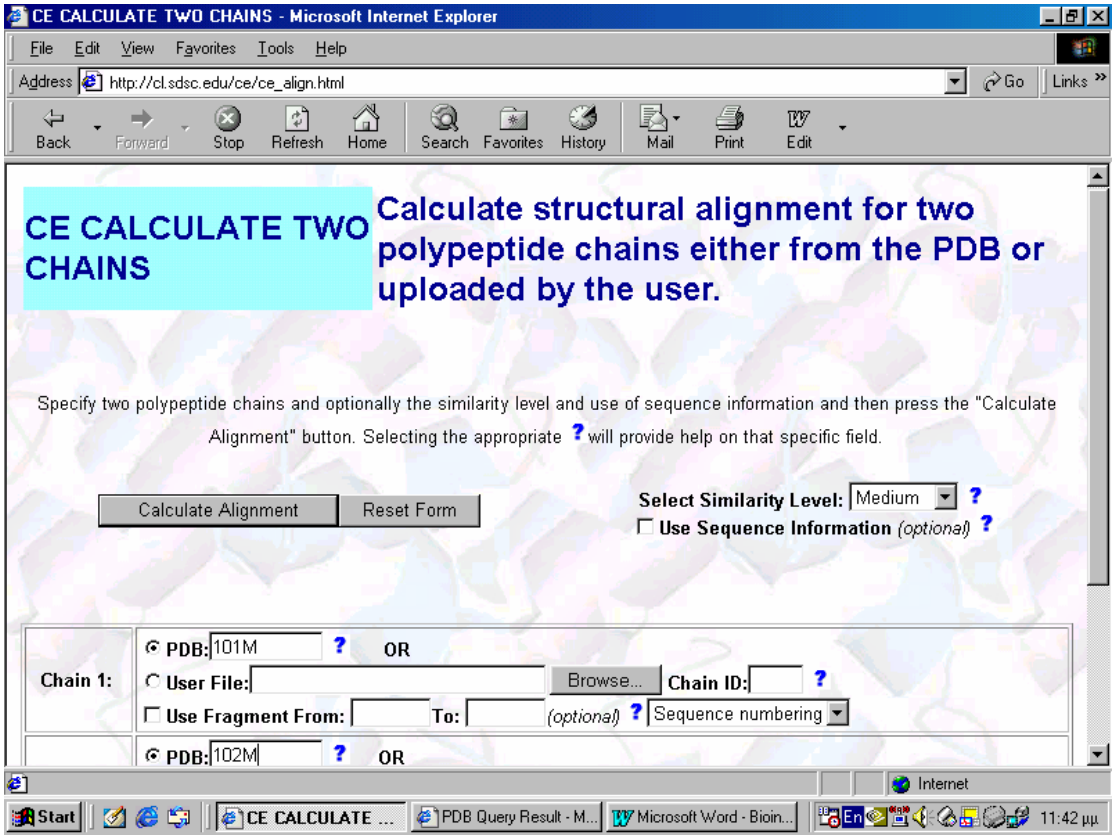


http://cl.sdsc.edu/ce.html.

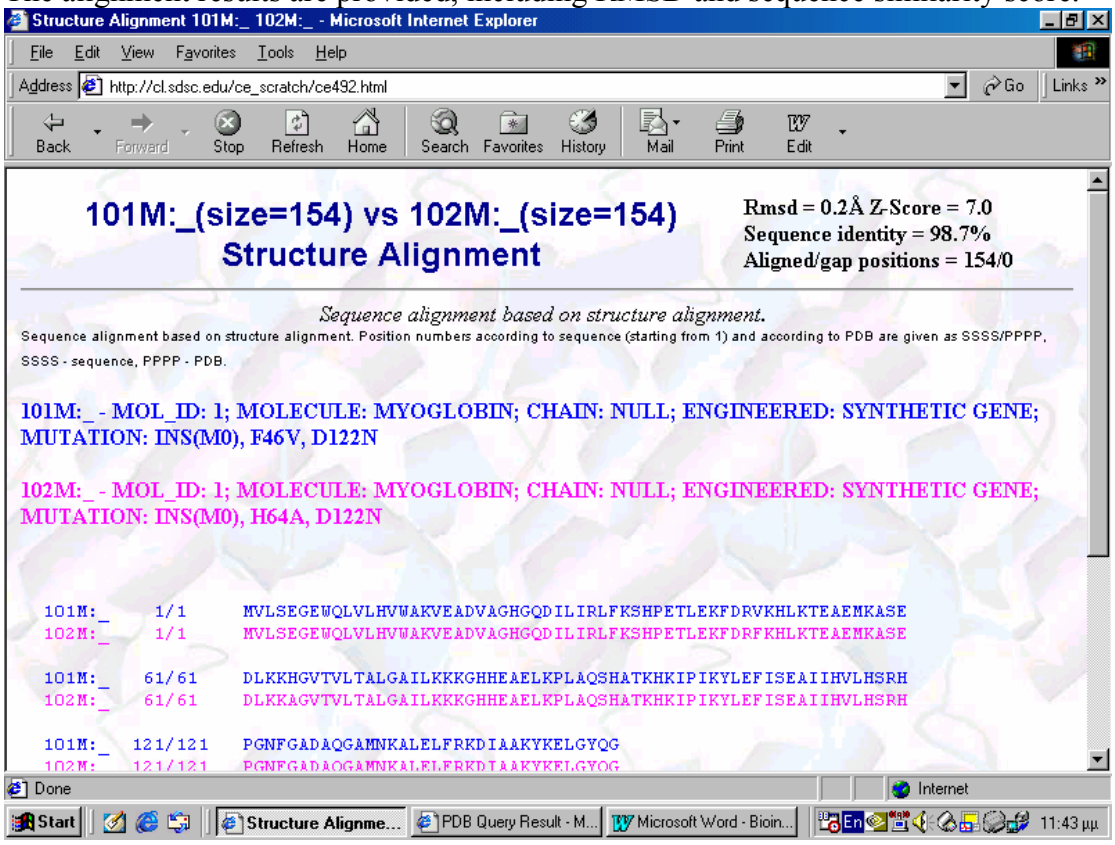
Στη συνέχεια πατήστε την επιλογή ‘TWO CHAINS’.

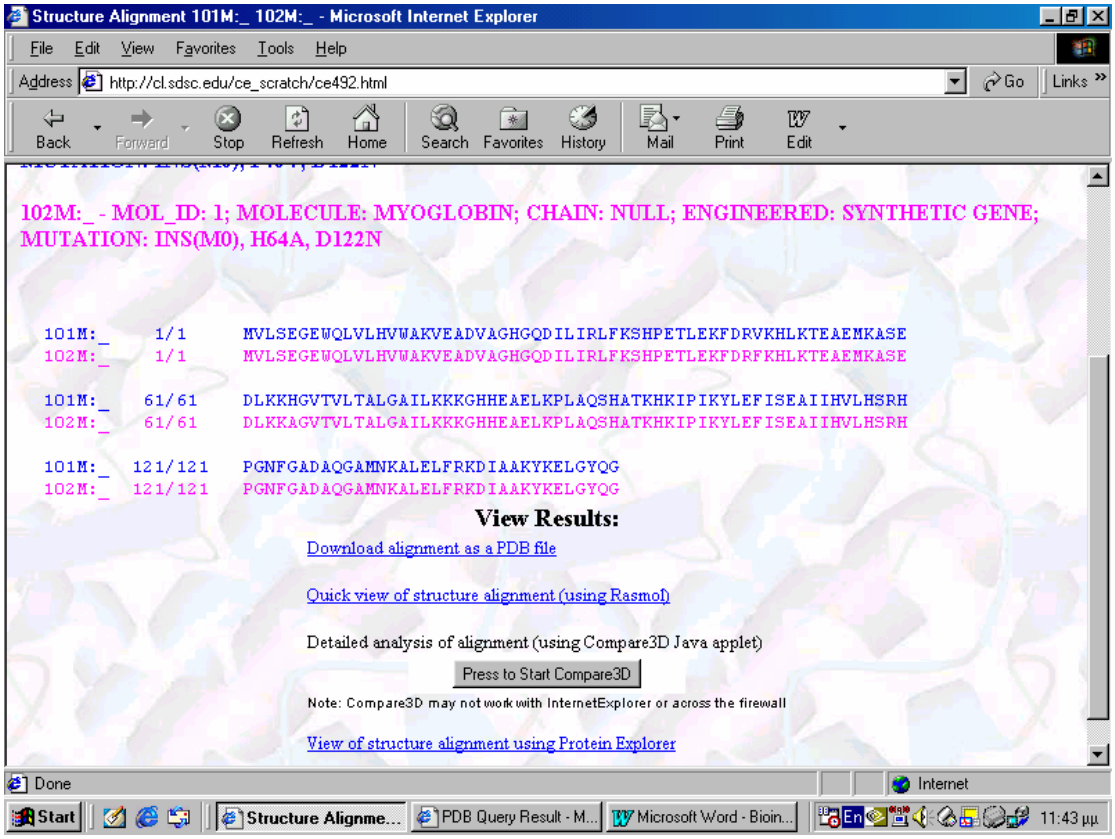


Για τη σύγκριση των πρωτεϊνών 101M και 102M, πληκτρολογήστε στο πεδίο της PDB για την Chain1: PDB: 101M και στο αντίστοιχο πεδίο για την Chain 2: PDB:102M. Έπειτα πατήστε το “Calculate Alignment’ και θα προκύψουν τα αποτελέσματα της αντιστοίχησης, συμπεριλαμβανομένου του RMSD και της βαθμολογίας για την ομοιότητα των αλληλουχιών.

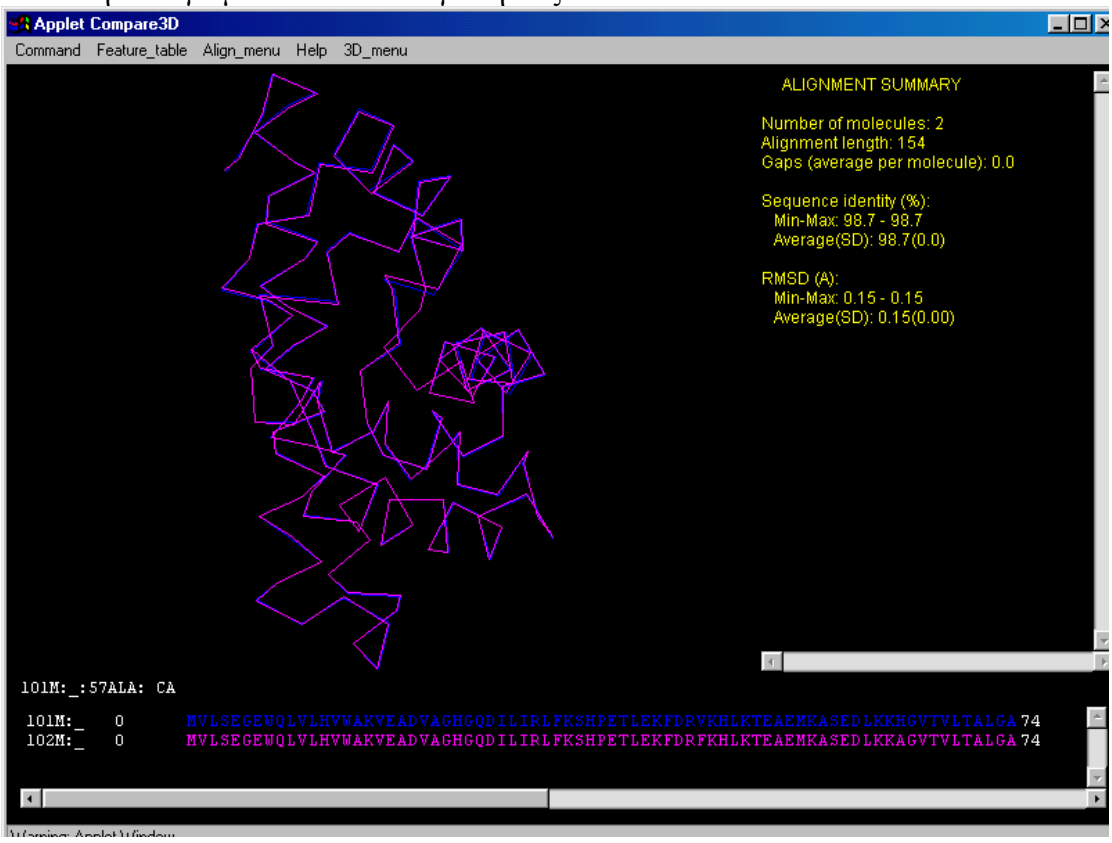


The alignment results are provided, including RMSD and sequence similarity score.

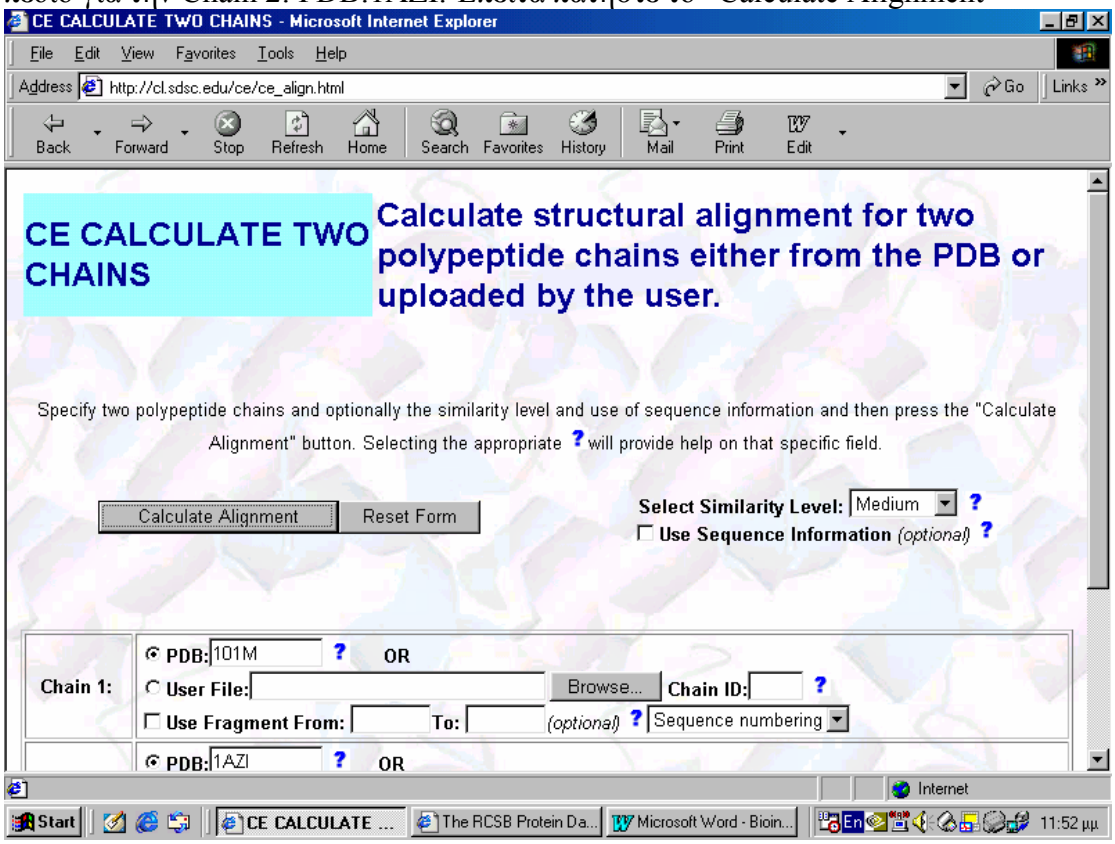




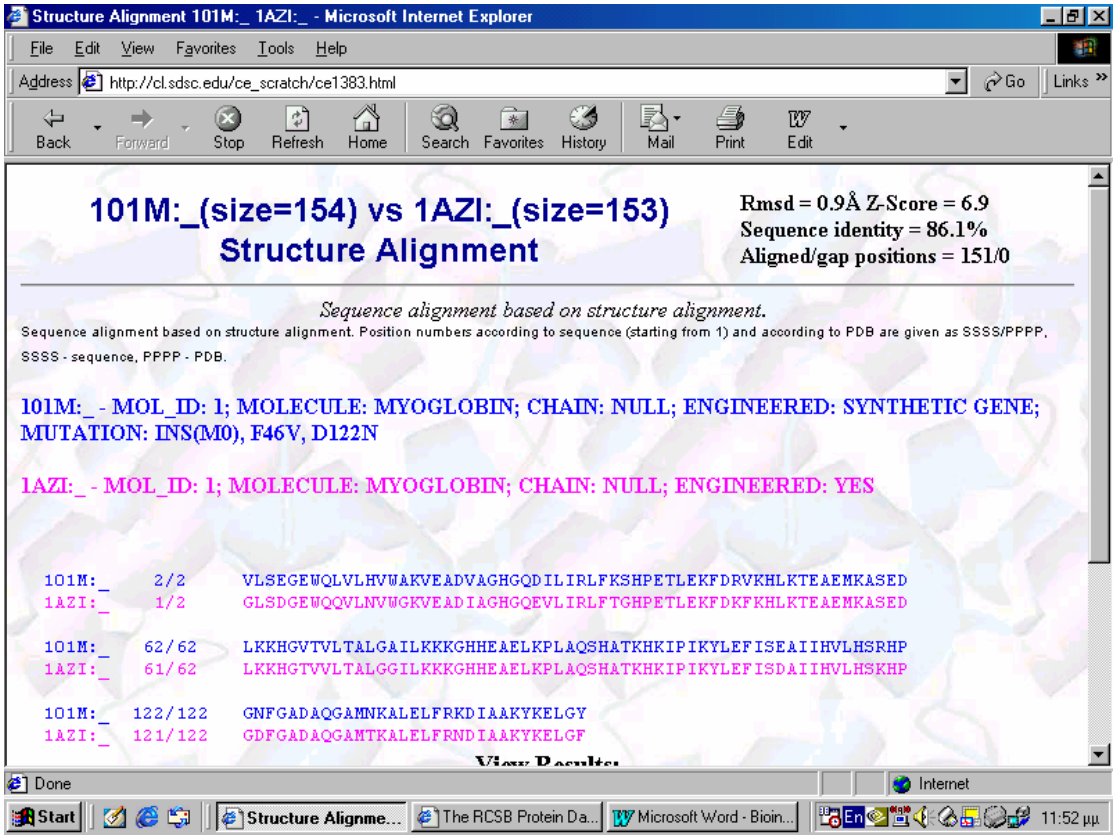
Πατώντας την επιλογή 'View of structure alignment using Protein Explorer' μπορείτε να δείτε τις δύο πρωτεΐνες, με την προϋπόθεση ότι υπάρχει εγκατεστημένο το κατάλληλο λογισμικό στον υπολογιστή σας.



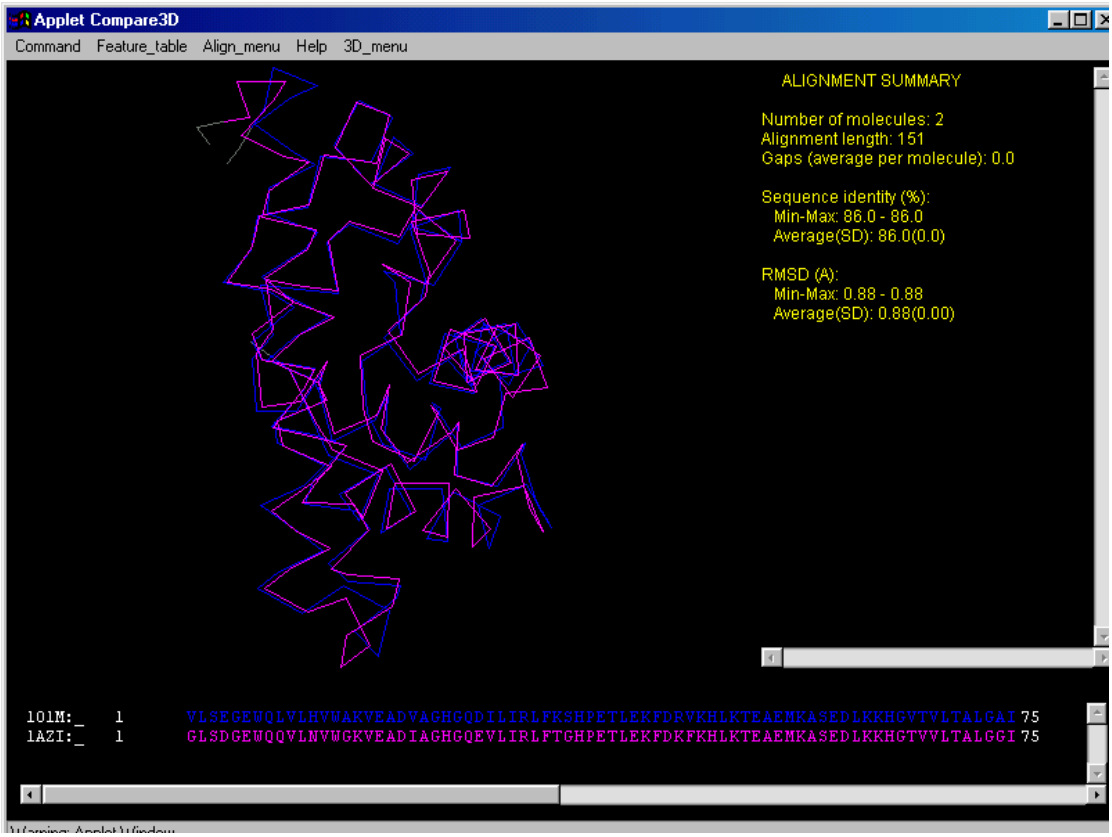
Για τη σύγκριση των πρωτεϊνών 101M και 1AZI (μία άλλη myoglobin), πληκτρολογήστε στο πεδίο της PDB για την Chain1: PDB: 101M και στο αντίστοιχο πεδίο για την Chain 2: PDB:1AZI. Έπειτα πατήστε το “Calculate Alignment’



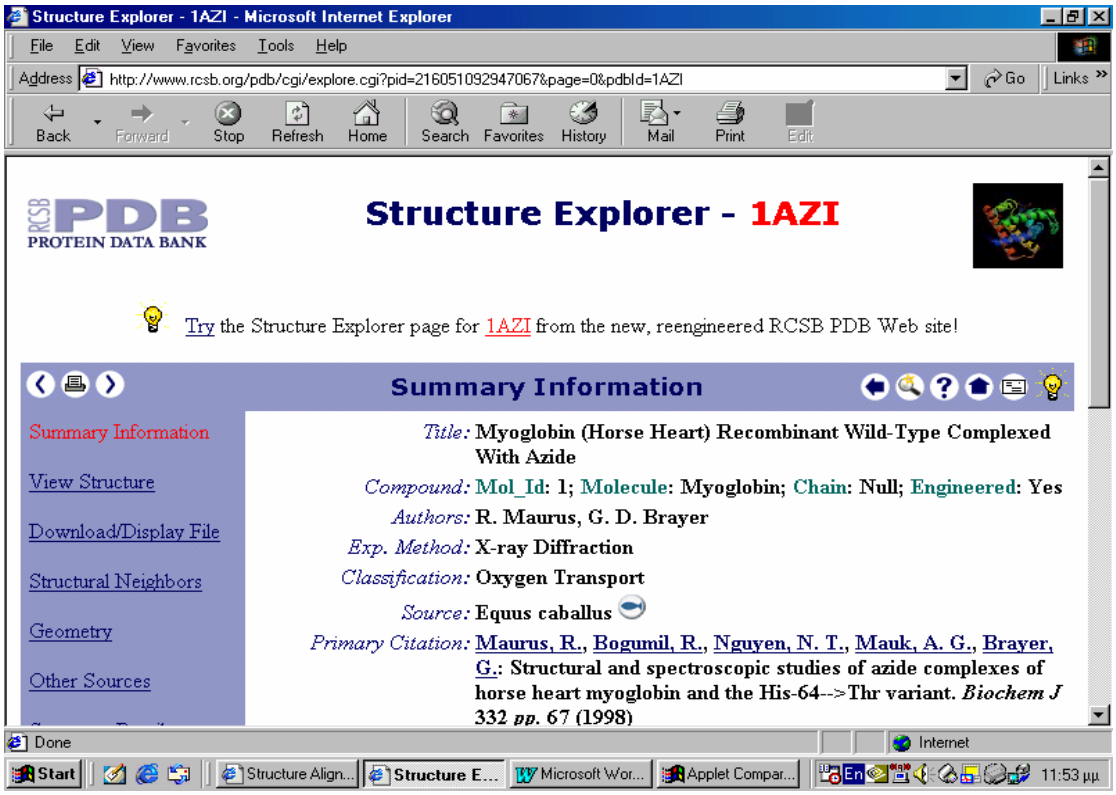
Τότε θα προκύψουν τα αποτελέσματα της αντιστοίχησης.



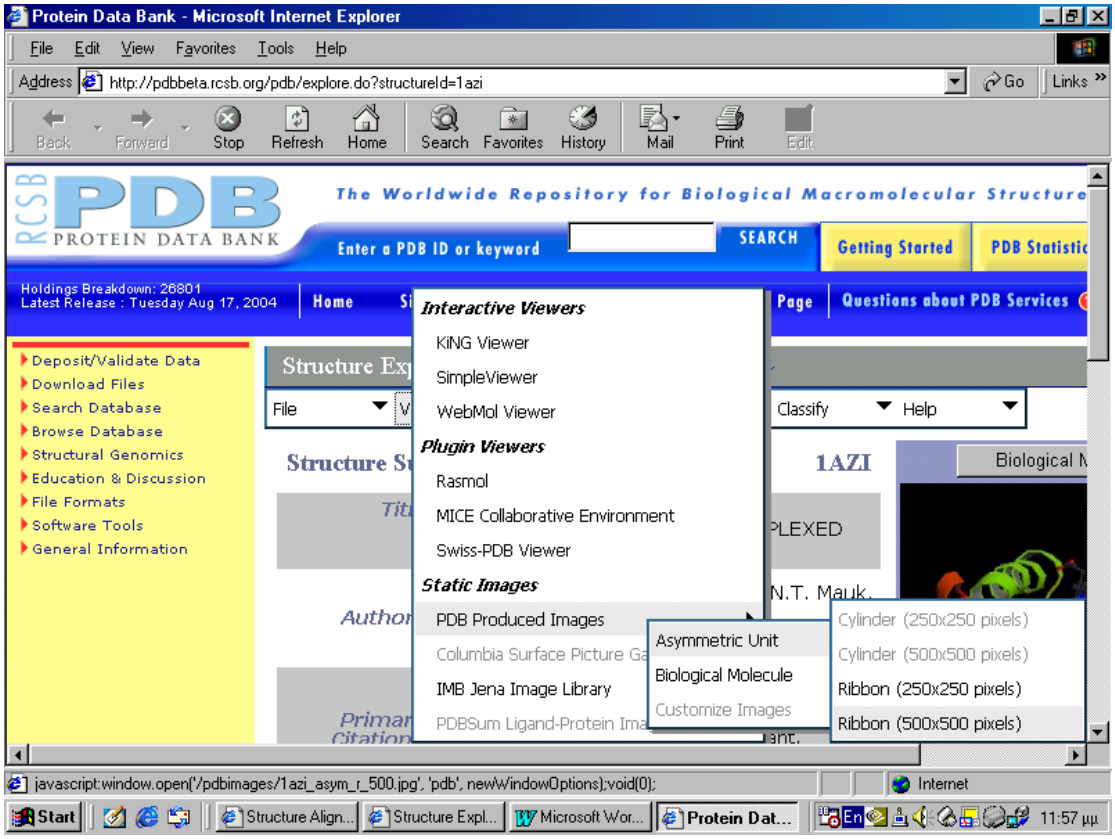
Και πάλι μπορείτε να δείτε τις δύο πρωτεΐνες πατώντας την επιλογή ‘View of structure alignment using Protein Explorer’.

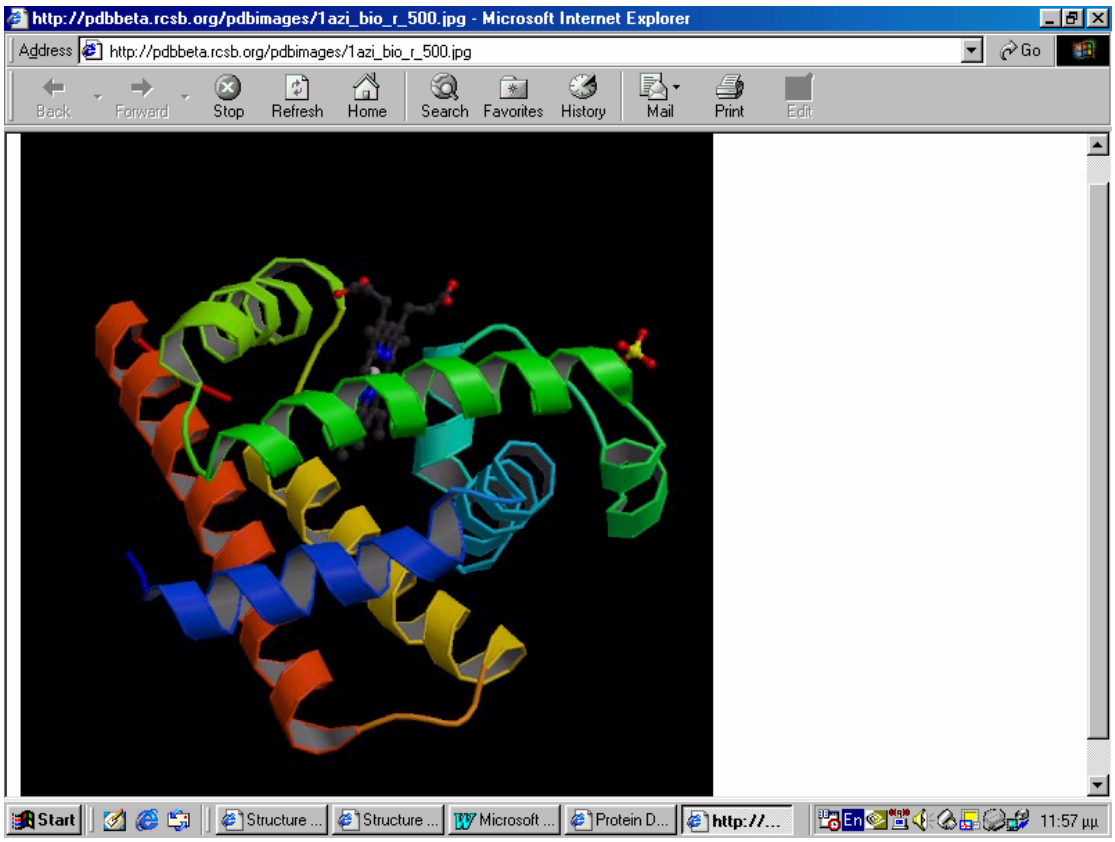


Επίσης, μπορούμε να βρούμε το μοντέλο με κορδέλες (ribbon model) για την 1AZI



μέσω της PDB, με τον τρόπο που ήδη δείξαμε..


Protein Data Bank - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://pdbbeta.rcsb.org/pdb/explore.do?structureId=1azi
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
RCSB PDB PROTEIN DATA BANK
The Worldwide Repository for Biological Macromolecular Structure
Enter a PDB ID or keyword
SEARCH
Getting Started
PDB Statistic
Holdings Breakdown: 26801
Latest Release : Tuesday Aug 17, 2004
Home
Page
Questions about PDB Services
Deposit/Validate Data
Download Files
Search Database
Browse Database
Structural Genomics
Education & Discussion
File Formats
Software Tools
General Information
Interactive Viewers
KiNG Viewer
SimpleViewer
WebMol Viewer
Plugin Viewers
Rasmol
MICE Collaborative Environment
Swiss-PDB Viewer
Static Images
PDB Produced Images
Columbia Surface Picture Ga
IMB Jena Image Library
PDBSum Ligand-Protein Ima
Asymmetric Unit
Biological Molecule
Customize Images
Cylinder (250x250 pixels)
Cylinder (500x500 pixels)
Ribbon (250x250 pixels)
Ribbon (500x500 pixels)
File
Classify
Help
1AZI
Biological M
N.T. Mauk.
javascript:window.open('/pdbimages/1azi_asym_r_500.jpg', 'pdb', newWindowOptions);void(0);
Internet
Start
Structure Align...
Structure Expl...
Microsoft Wor...
Protein Dat...
11:57 µµ


http://pdbbeta.rcsb.org/pdbimages/1azi_bio_r_500.jpg - Microsoft Internet Explorer
Address http://pdbbeta.rcsb.org/pdbimages/1azi_bio_r_500.jpg
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Start
Structure ...
Structure ...
Microsoft ...
Protein D...
http://...
11:57 µµ

Βάσεις δεδομένων δομών (Structure databases)

Πολλές πρωτεΐνες παρουσιάζουν ομοιότητες στη δομή τους, και οι ομοιότητες αυτές μπορεί να μαρτυρούν κοινή εξελικτική προέλευση. Η εξελικτική πορεία περιλαμβάνει αλλαγές στην αλληλουχία των αμινοξέων, όπως είναι οι αντικαταστάσεις, οι εισαγωγές και οι διαγραφές. Σε περιπτώσεις πρωτεϊνών με μακρινή συγγένεια, οι αλλαγές αυτές έχουν προκαλέσει «διπλώσεις», στις οποίες το πλήθος και η κατεύθυνση των δευτεροταγών δομών ποικίλει .

Πολλά πλάνα κατηγοριοποίησης έχουν προταθεί στην προσπάθεια για καλύτερη κατανόηση των σχέσεων δομών / αλληλουχιών και της εξελικτικής πορείας η οποία δημιουργεί διαφορετικές οικογένειες «διπλώσεων». Τα πλάνα αυτά διαφοροποιούνται ως προς τις μεθόδους που χρησιμοποιούν για την αναγνώριση και την αξιολόγηση της δομικής ομοιότητας.

Οι οικογένειες δομών δημιουργούνται με δύο μεθόδους: 1) χρησιμοποιώντας αλγόριθμους οι οποίοι αναζητούν και ομαδοποιούν με βάση κοινά πρότυπα, 2) χρησιμοποιώντας διαδικασίες οι οποίες βασίζονται σε καθολική σύγκριση δομών, οπτική ή μαθηματική.

Η πλέον διαδεδομένη μέθοδος κατηγοριοποίησης είναι η CATH, η οποία βρίσκεται ενσωματωμένη σε μία βάση δεδομένων δομών.

CATH

Η βάση δεδομένων CATH (Class, Architecture, Topology, Homology) είναι μία ιεραρχική κατηγοριοποίηση περιοχών των δομών των πρωτεϊνών. Η CATH συντηρείται στο UCL. Η κατηγοριοποίηση βασίζεται σε καθολική, μαθηματική και οπτική, σύγκριση δομών. Οι διάφορες κατηγορίες στο πλάνο κατηγοριοποίησης αναγνωρίζονται με τη βοήθεια μοναδικών αριθμών και περιγραφικών ονομάτων.

Τα επίπεδα ιεραρχίας της δομικής κατηγοριοποίησης είναι:

Κατηγορία πρωτεϊνών (Protein class) : Η κατηγορία μίας δομής πρωτεϊνών αντιπροσωπεύει το ποσοστό των α-ελίκων (α-helices) ή των β-φύλλων (β-strands) μέσα σε μία τρισδιάστατη δομή. Οι βασικές κατηγορίες είναι: κυρίως-α (mainly-α), κυρίως-β (mainly-β), εναλλαγές α/β(alternating α/β) και α+β (μείξη α-β), καθώς και οι πρωτεΐνες εκείνες με χαμηλή περιεκτικότητα σε δευτεροταγείς δομές.

Αρχιτεκτονική πρωτεϊνών (Protein architecture) : Εδώ γίνεται περιγραφή της γενικής κατανομής των δευτεροταγών δομών (α-helices και β-strands) σε τρισδιάστατο χώρο, χωρίς να λαμβάνονται υπ' όψιν οι δεσμοί τους. Η καταχώρηση μίας πρωτεΐνη γίνεται με το χέρι και χρησιμοποιώντας απλές περιγραφές για την κατανομή των δευτεροταγών δομών (π.χ. barrel, roll, sandwich, κ.λ.π.).

Τοπολογία πρωτεϊνών (Protein topology) : Εδώ γίνεται περιγραφή της γενικής κατανομής των δευτεροταγών δομών σε τρισδιάστατο χώρο, λαμβάνοντας όμως υπ' όψιν τον προσανατολισμό των δευτεροταγών δομών και τους μεταξύ τους δεσμούς. Οι περιοχές ομαδοποιούνται χρησιμοποιώντας παραμέτρους οι οποίες έχουν προκύψει εμπειρικά , μέσω αλγορίθμων σύγκρισης δομών. Πρωτεΐνες οι οποίες παρουσιάζουν δομική ομοιότητα >60% καταχωρούνται στην ίδια τοπολογία.

Ομόλογη υπέρ-οικογένεια (Homologous superfamily) : Μία ομόλογη υπερ-οικογένεια είναι μία ομάδα πρωτεϊνών των οποίων οι δομές δείχνουν ότι υπάρχει μία κοινή εξελικτική προέλευση (δηλαδή οι πρωτεΐνες αυτές είναι ομόλογες). Οι

ομοιότητες εντοπίζονται πρωτίστως μέσω της σύγκρισης αλληλουχιών και εν συνεχεία μέσω αλγορίθμου σύγκρισης δομών. Περιοχές οι οποίες παρουσιάζουν ομοιότητα αλληλουχιών >35% τοποθετούνται στην ίδια ομάδα.

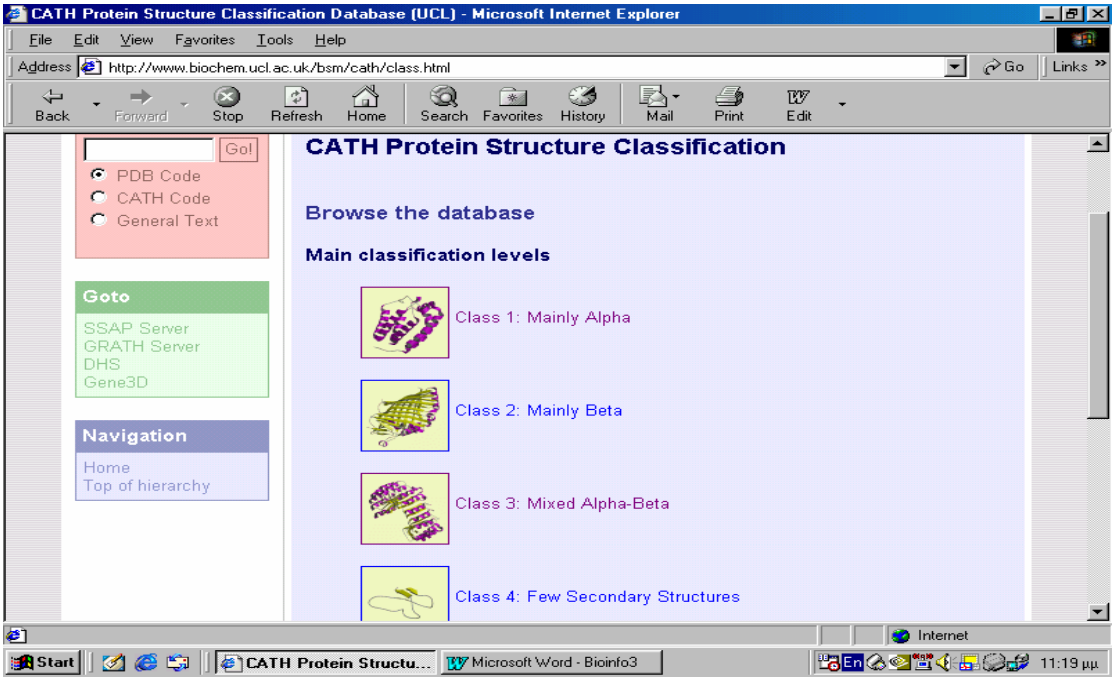
Οικογένεια αλληλουχιών (Sequence family) : Δομές οι οποίες έχουν ήδη διαχωριστεί μέσα σε ομάδες ομολογίας, μπορούν να διαχωριστούν περαιτέρω με βάση την ομοιότητα των αλληλουχιών τους. Οι πρωτεΐνες οι οποίες έχουν διαχωριστεί σε οικογένειες προφανώς σχετίζονται μέσω της εξέλιξης. Αυτό σημαίνει ότι ανάμεσα στις πρωτεΐνες, οι ομοιότητες των residues ανά ζεύγη είναι >35%. Σε ορισμένες περιπτώσεις, και ελλείψει υψηλής ομοιότητας αλληλουχιών, παρόμοιες λειτουργίες και δομές προσφέρουν αποδείξεις για την κοινή προέλευση των πρωτεϊνών.

ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

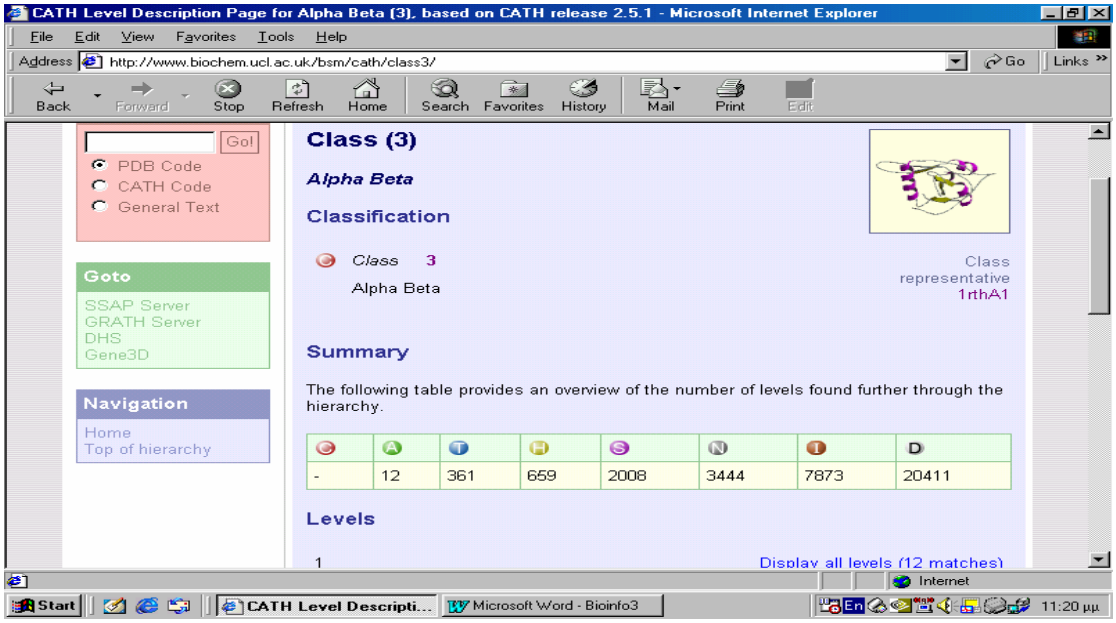
CATH – Κατηγοριοποίηση δομών (Structure classification)

Για να επισκεφθείτε την ιστοσελίδα της CATH, ανοίξτε τον Internet Explorer και πληκτρολογήστε τη διεύθυνση www.biochem.ucl.ac.uk/bsm/cath/. Στη συνέχεια, για να δείτε την ιεραρχική κατηγοριοποίηση συγκεκριμένων πρωτεϊνών, πατήστε στην επιλογή "Browse or search the classification".

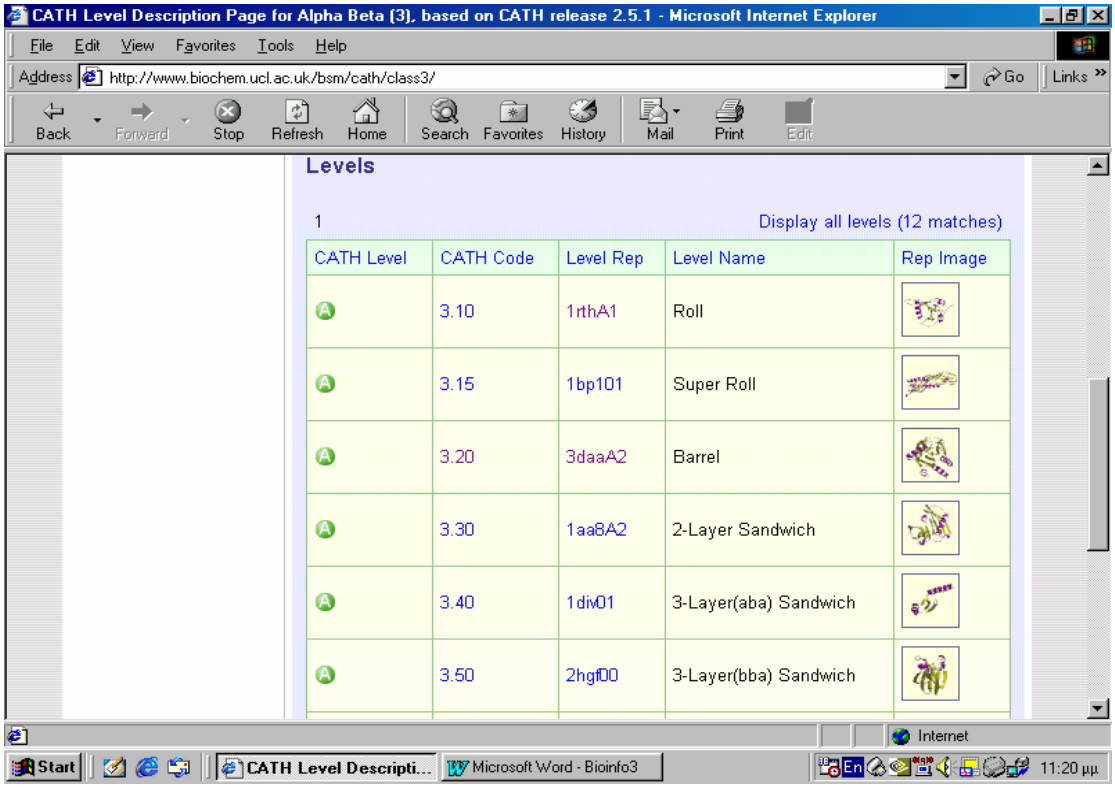
Στη νέα σελίδα παρουσιάζονται τα 4 διαφορετικά επίπεδα κατηγοριών. Προκειμένου να δείτε την κατηγορία α+β (μείξη α-β) πατήστε στο "Class 3: Mixed Alpha-Beta".



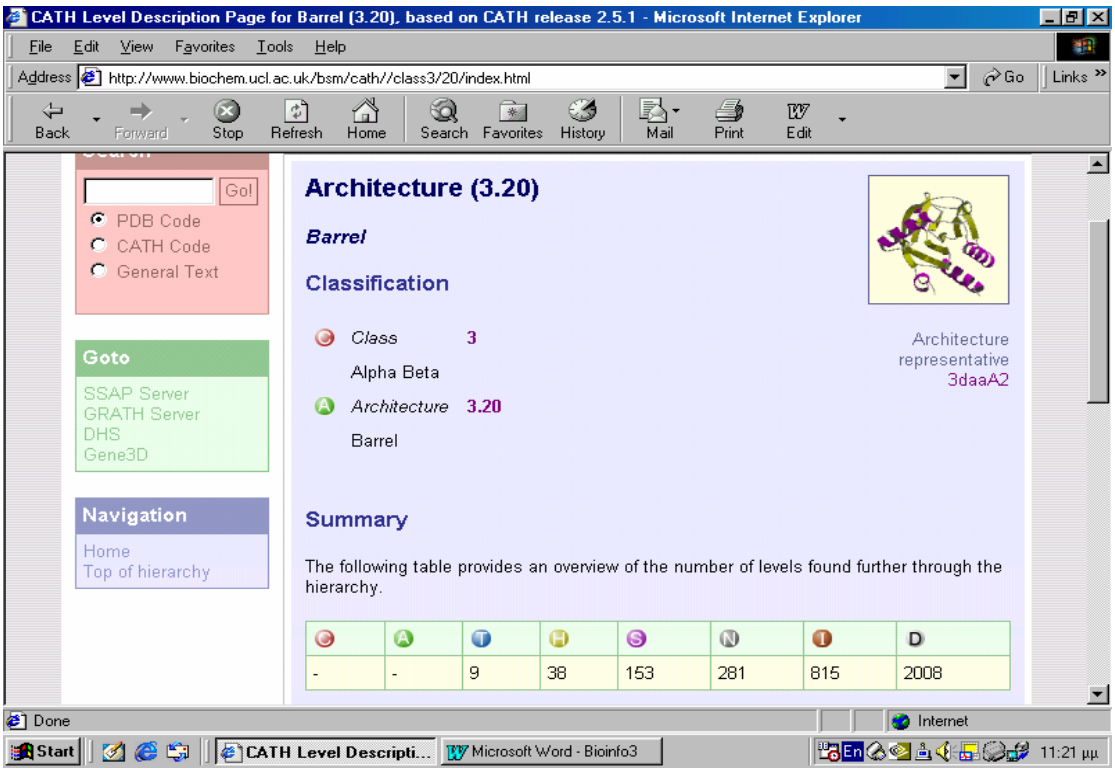
Μπορείτε να δείτε ένα συγκεντρωτικό πίνακα με το πλήθος των εγγραφών για κάθε ένα από τα επίπεδα τα οποία συναντώνται μέσα στην ιεραρχία. Η σειρά των επιπέδων είναι όπως αναφέρθηκε προηγουμένως και το κάθε επίπεδο συμβολίζεται με το αρχικό γράμμα του ονόματός του (A για το Architecture, κ.ο.κ.). Επίσης παρατίθεται ένας πίνακας με τα επίπεδα αρχιτεκτονικής τα οποία συναντώνται μέσα στην τρίτη κατηγορία. Για την τρίτη κατηγορία (class 3) υπάρχουν 12 επίπεδα αρχιτεκτονικής. Πατώντας στο A βλέπετε τα 12 αυτά επίπεδα.


CATH Level Description Page for Alpha Beta (3), based on CATH release 2.5.1 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.biochem.ucl.ac.uk/bsm/cath/class3/
Go
Links
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Go!
PDB Code
CATH Code
General Text
Goto
SSAP Server
GRATH Server
DHS
Gene3D
Navigation
Home
Top of hierarchy
Class (3)
Alpha Beta
Classification
Class 3
Alpha Beta
Class representative 1rthA1
Summary
The following table provides an overview of the number of levels found further through the hierarchy.
C A T H S N I D
- 12 361 659 2008 3444 7873 20411
Levels
1
Display all levels (12 matches)
Internet
Start
CATH Level Descripti...
Microsoft Word - Bioinfo3
11:20 µµ

Εάν τώρα επιλέξετε κάποιο επίπεδο και πατήσετε στον αριθμό που αντιστοιχεί στο CATH Code αυτού, θα οδηγηθείτε στην επόμενη κατηγορία που είναι η τοπολογία και όπου παραθέτονται οι πρωτεΐνες οι οποίες ταξινομούνται μαζί με αυτή την οποία επιλέξατε με τα τοπολογικά αυτή τη φορά κριτήρια.
Παραδείγματος χάριν, αν επιλέξετε το επίπεδο 'Barrel' και πατήσετε στο 3.20 (CATH Code), τότε θα οδηγηθείτε σε μία σελίδα όπου παρουσιάζονται τα 9 επίπεδα τοπολογίας που αντιστοιχούν στην αρχιτεκτονική κατηγορία 'Barrel'.

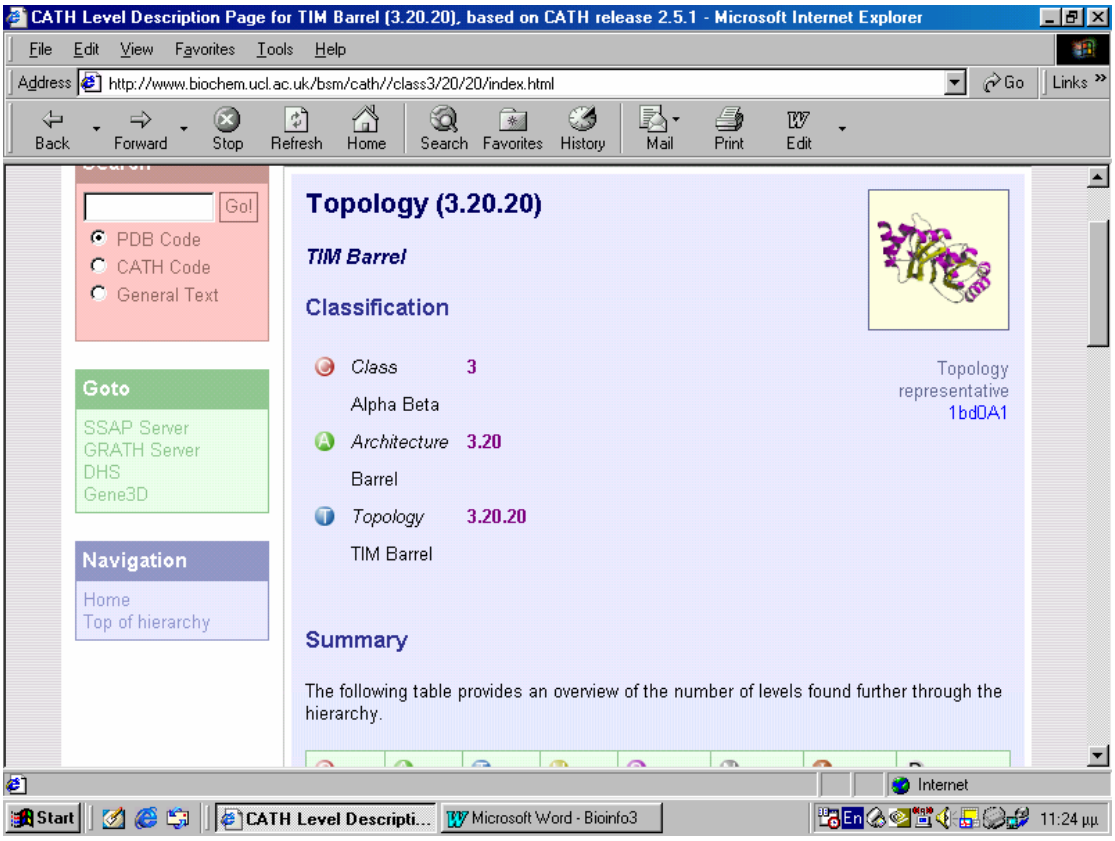


Πατώντας στο γράμμα T μπορείτε να δείτε τα 9 αυτά επίπεδα τοπολογίας.

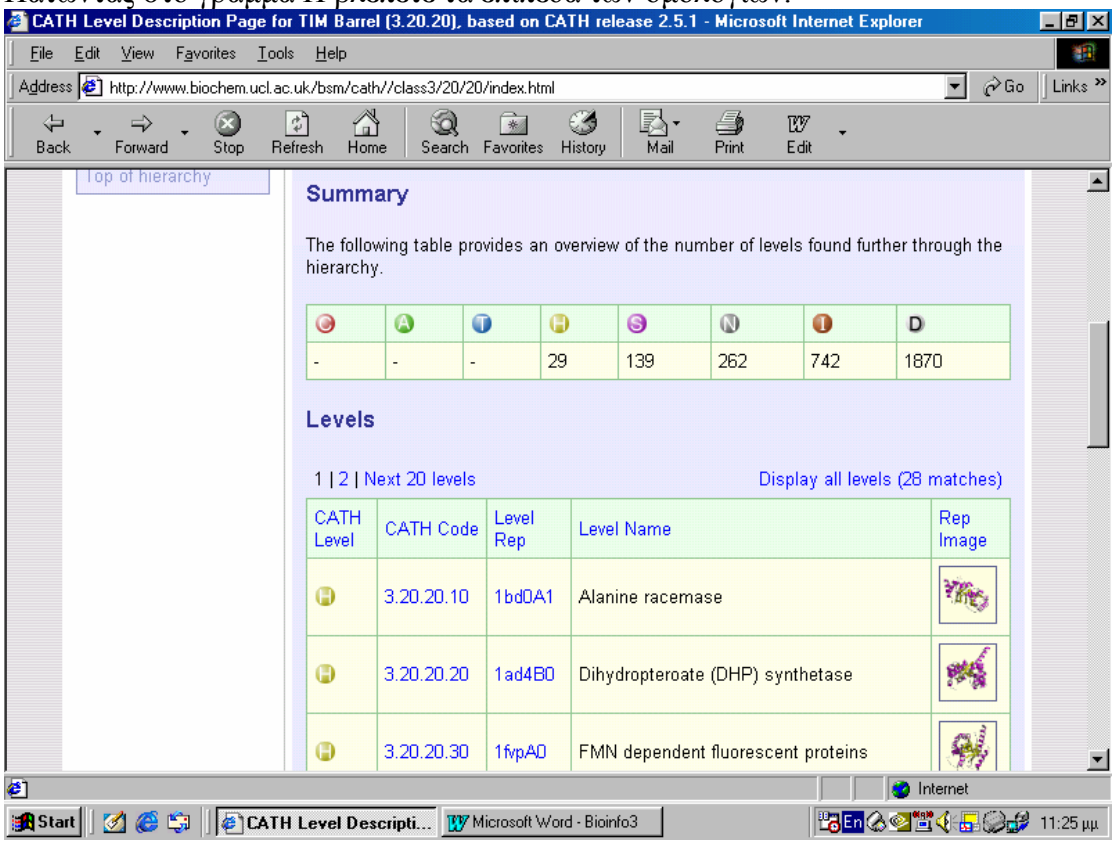


Τώρα, πατήστε στο 3.20.20, TIM Barrel για να δείτε το επόμενο επίπεδο της ιεραρχίας.

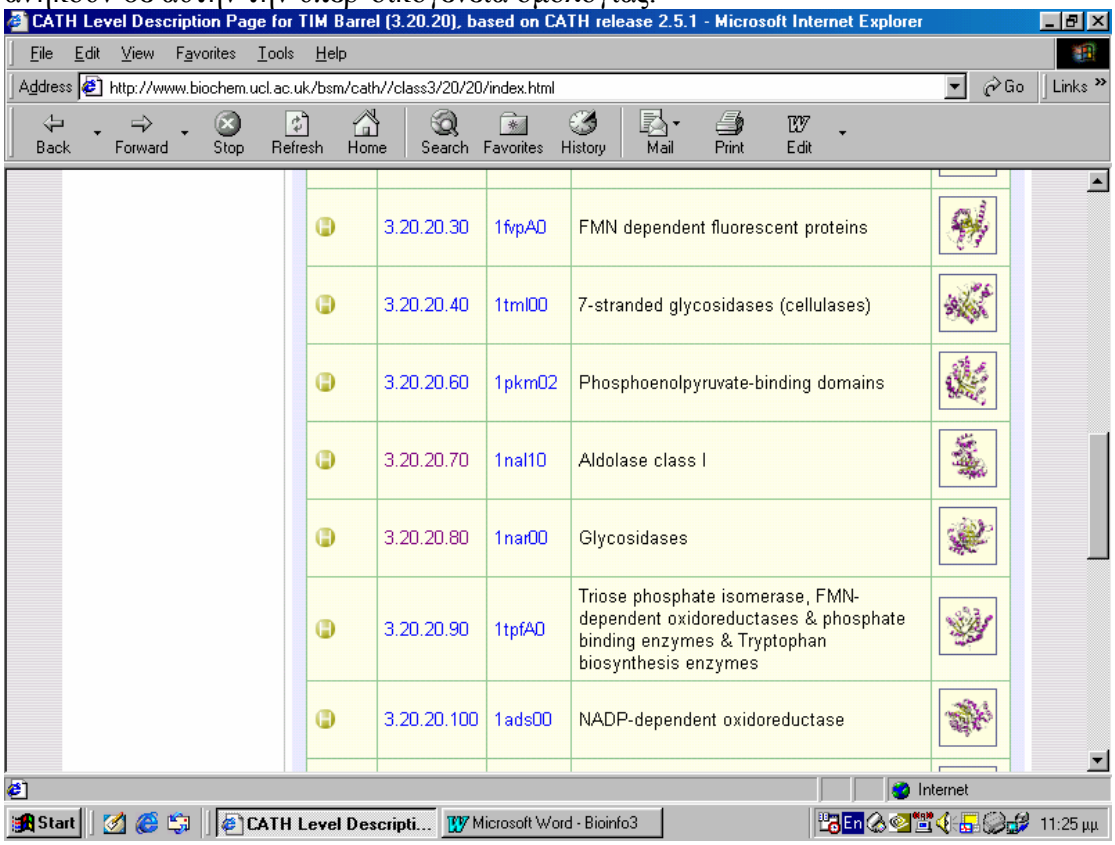
Στο TIM Barrel αντιστοιχούν 29 ομολογίες (Homologies).



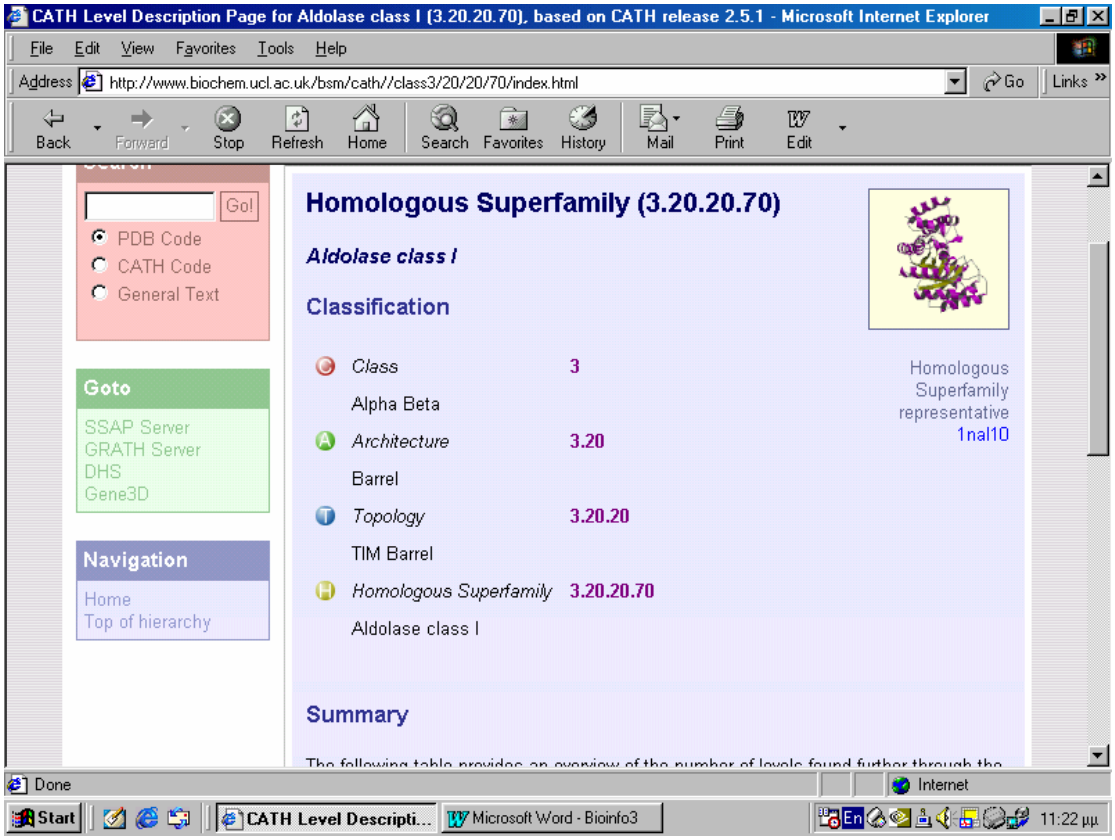
Πατώντας στο γράμμα Η βλέπετε τα επίπεδα των ομολογιών.



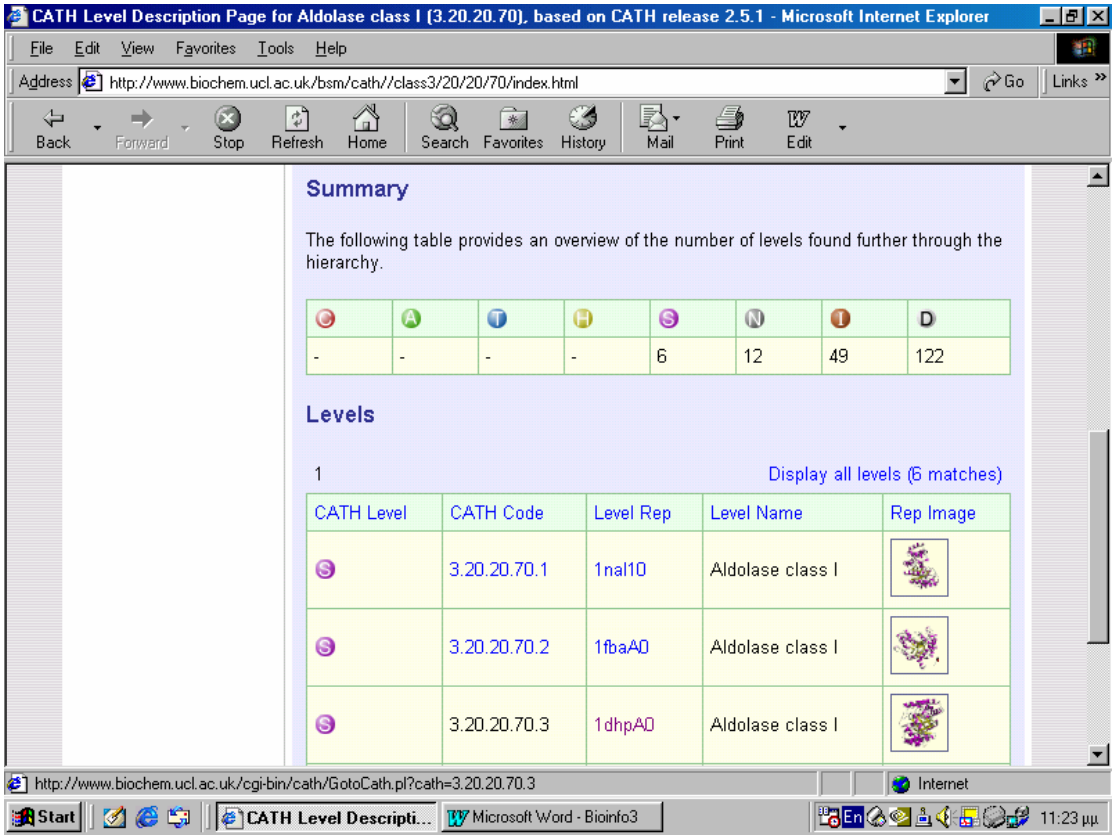
Πατήστε την 3.20.20.70, Aldolase κατηγορία I για να δείτε τις αλληλουχίες οι οποίες ανήκουν σε αυτήν την υπερ-οικογένεια ομολογίας.



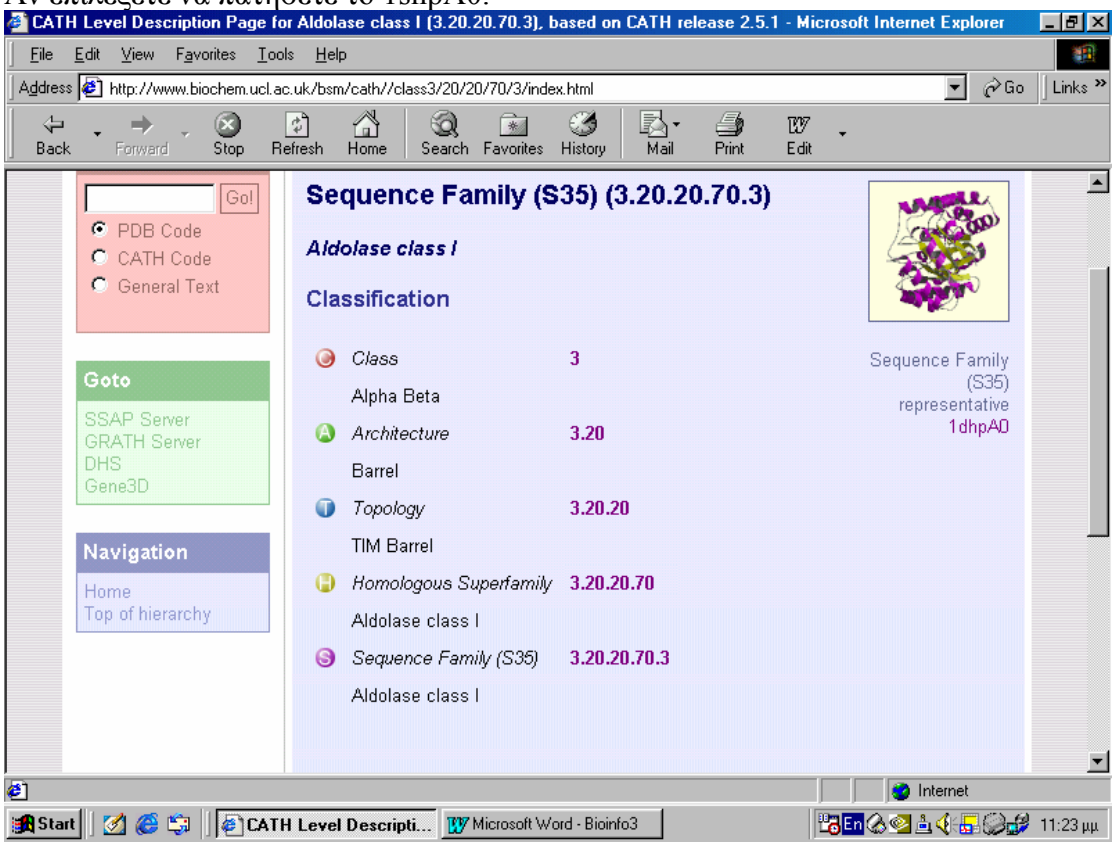
Υπάρχουν 6 οικογένειες αλληλουχιών οι οποίες ανήκουν στην υπέρ-οικογένεια ομολογίας της Aldolase κατηγορία I.

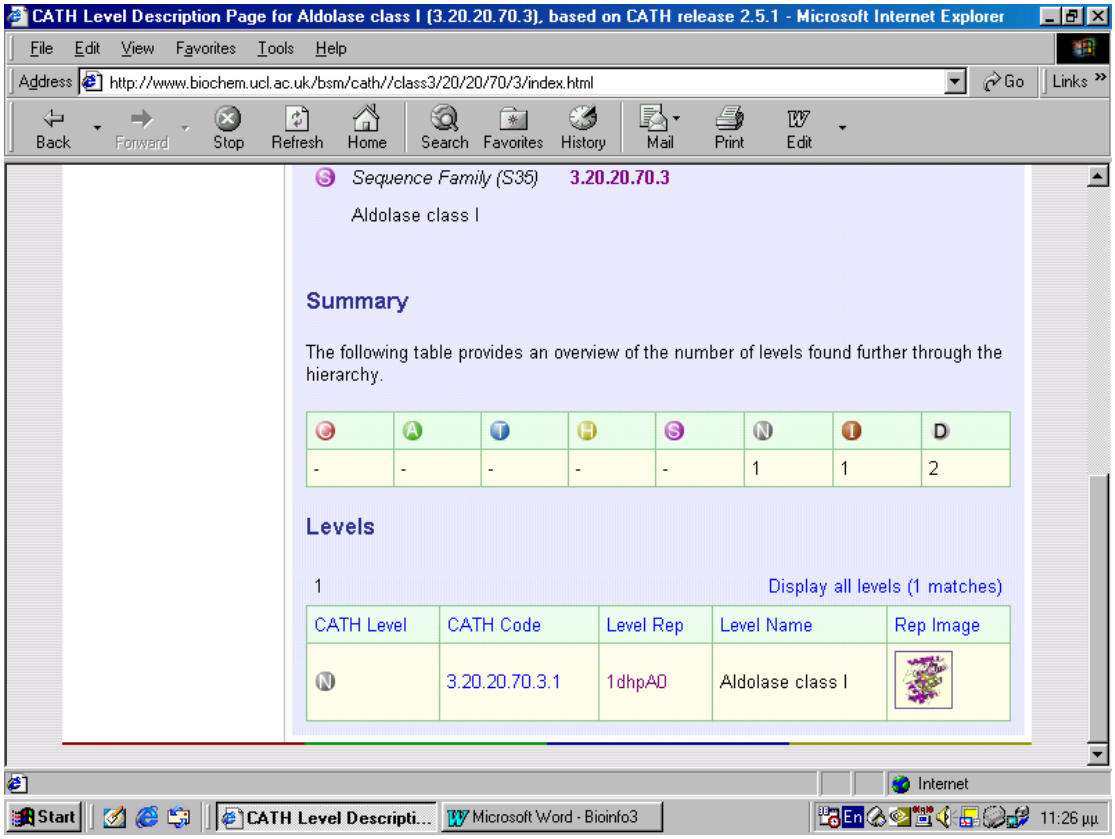


Πατήστε στο γράμμα S για να δείτε τις 6 οικογένειες αλληλουχιών.

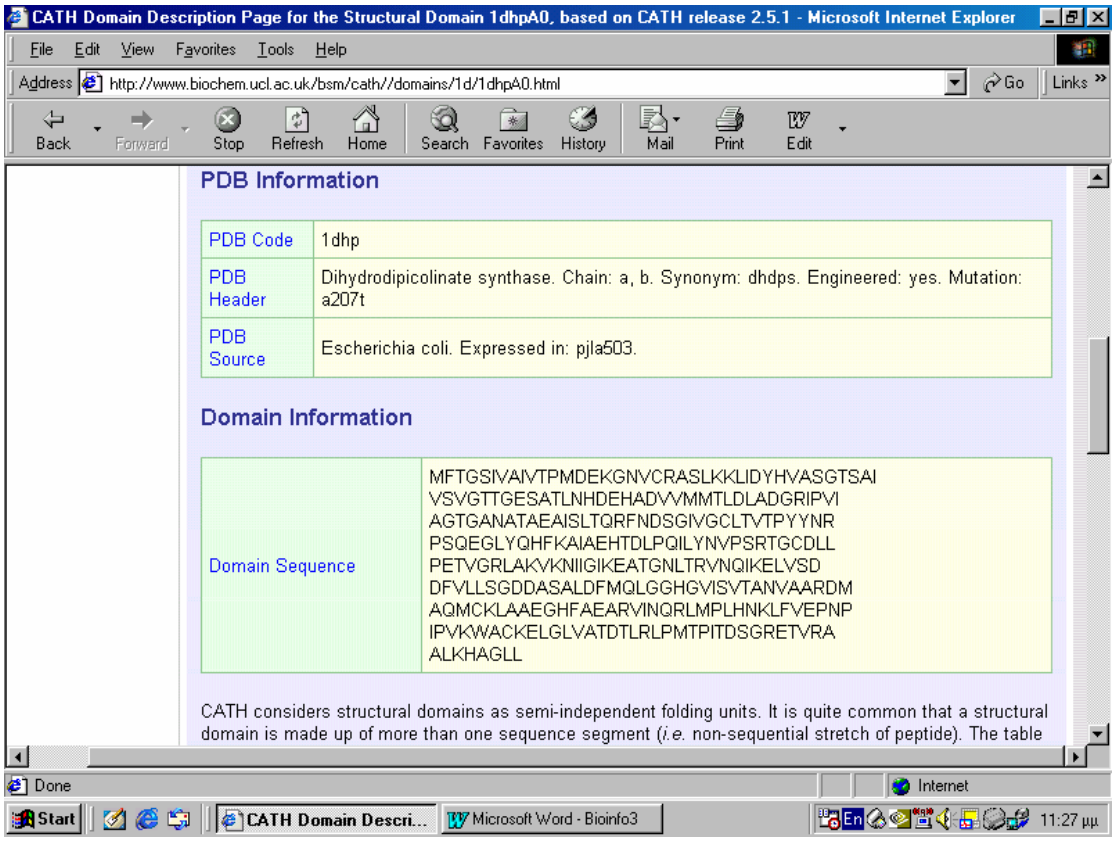


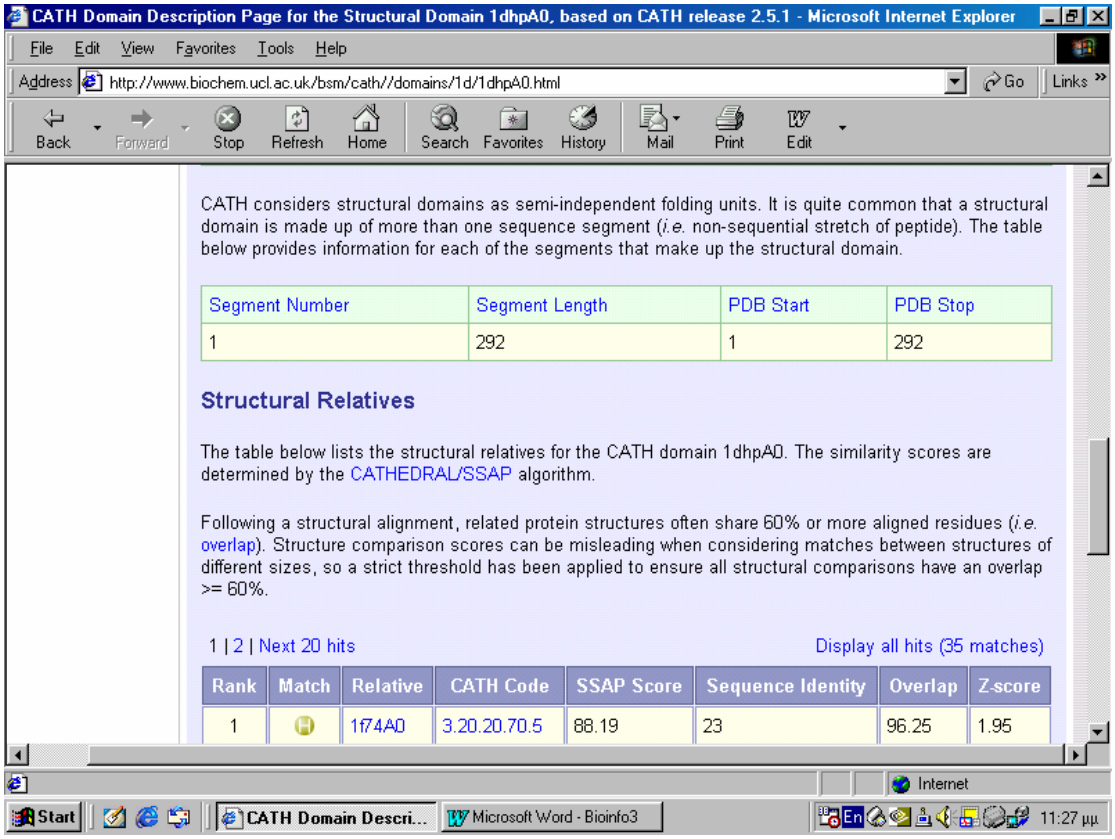
Επομένως, η 1dhpA0 ανήκει στην Aldolase κατηγορία I, είναι TIM Barrel και αποτελεί μείξη a-b.
Αν επιλέξετε να πατήσετε το 1shpA0.



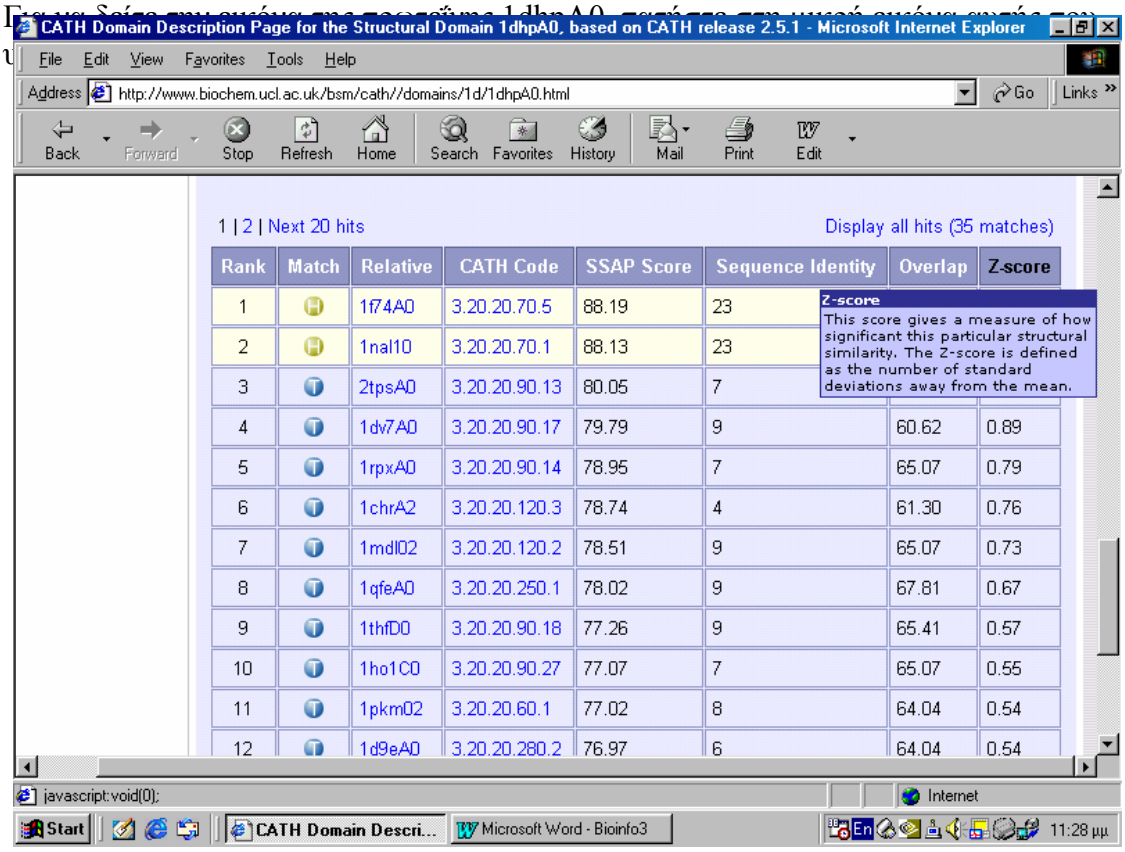


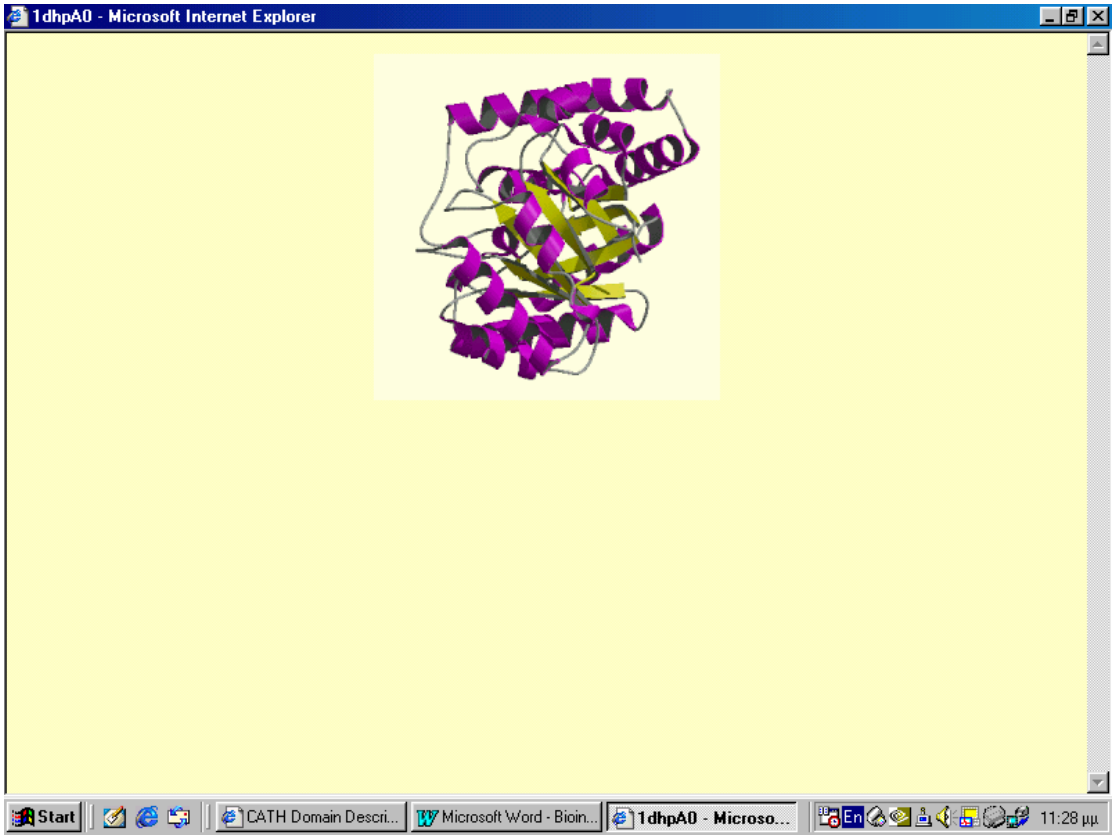
Τότε εμφανίζεται η αλληλουχία της πρωτεΐνης και σχετικές με αυτήν πληροφορίες.





Για να δούμε την ομάδα στην οποία ανήκει η πρωτεΐνη 1dhpA0, κατεβαίνουμε στην κορυφή σελίδας και υ




1dhpA0 - Microsoft Internet Explorer
Start
CATH Domain Descri...
Microsoft Word - Bioin...
1dhpA0 - Microso...
11:28 μμ

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Alberts P. (2002). Molecular Biology of the Cell
- Armstrong D.J. (2004). Lecture notes Bioinformatics 2. University of Edinburgh
- Durbin R., Eddy S., Krogh A. & Mitchinson G. (1998) Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press
- Orengo et al. (2003) Bioinformatics. BIOS
- Attwood and Parry-Smith (1999) Introduction to bioinformatics. Prentice Hall
- Westhead et al. (2002) Bioiformatics. BIOS
- Mount (2001) Bioinfrmatics. CSHL Press.
- Jones and Karpp (1986) Introducing Genetics. Murray