Βιοπληροφορική

Έρευνα βάσεων δεδομένων

Τόσο ο αλγόριθμος των Needleman & Wunsch, όσο και των Smith & Waterman ανήκουν στην κατηγορία των λεγόμενων 'σχολαστικών' (rigorous) αλγόριθμων στοίχισης, είναι δηλαδή αλγόριθμοι οι οποίοι εγγυούνται ότι η βέλτιστη στοίχιση μεταξύ των αλληλουχιών (ολική ή τοπική) θα βρεθεί. Το υπολογιστικό κόστος της πραγματοποίησης μιας στοίχισης με αυτούς τους αλγόριθμους είναι ανάλογο του γινομένου των μηκών των αλληλουχιών.

Έρευνα βάσεων δεδομένων

Αν και όπως έχει ήδη αναφερθεί, ένα υπολογιστικό κόστος ανάλογο του γινομένου των μηκών των αλληλουχιών αντιπροσωπεύει μια τεράστια βελτίωση σε σχέση με ένα εκθετικό κόστος, για την περίπτωση της έρευνας των βάσεων δεδομένων (με εκατοντάδες χιλιάδες ή και εκατομμύρια (ESTs) αλληλουχίες), αυτοί οι αλγόριθμοι είναι μάλλον μη ικανοποιητικοί από άποψη ταχύτητας. Για αυτό το λόγο έχουν αναπτυχθεί οι λεγόμενοι ευρεστικοί αλγόριθμοι.

Ευρεστικοί αλγόριθμοι

Οι ευρεστικοί αλγόριθμοι χρησιμοποιούν προσεγγίσεις οι οποίες επιτρέπουν την αναζήτηση ομολόγων πρωτεϊνών στις βάσεις δεδομένων να γίνει πολύ πιο γρήγορα (απ'ότι μέσω των σχολαστικών αλγόριθμων). Η χρήση αυτών των προσεγγίσεων έχει ως συνέπεια το ότι οι ευρεστικοί αλγόριθμοι δεν εγγυώνται ότι όλες οι ομόλογες πρωτεΐνες θα βρεθούν κατά την έρευνα, αν και η πιθανότητα να διαφύγουν στενά συσχετιζόμενες αλληλουχίες είναι μάλλον μικρή. Οι αλγόριθμοι αυτοί είναι τόσο πιο γρήγοροι από τους σχολαστικούς αλγόριθμους ώστε να τους έχουν σχεδόν πλήρως υποκαταστήσει για την έρευνα των βάσεων δεδομένων.

Ευρεστικοί αλγόριθμοι

Οι δύο πλέον γνωστοί ευρεστικοί αλγόριθμοι είναι αυτοί που κωδικοποιούνται στα προγράμματα FASTA και BLAST (Basic Local Alignment Search Tool).
Η θεμελιώδης προσέγγιση που χρησιμοποιούν και οι δύο αλγόριθμοι για να επιταχύνουν την διαδικασία έρευνας στηρίζεται στη διάσπαση των αλληλουχιών σε 'λέξεις', δηλαδή σε μικρού μήκους υπο-ακολουθίες διαδοχικών (στις αλληλουχίες) καταλοίπων. Η υπόθεση που χρησιμοποιείται είναι ότι δύο ομόλογες αλληλουχίες θα έχουν τουλάχιστον μια 'λέξη' η οποία θα είναι κοινή (παρούσα) και στις δύο.

Ευρεστικοί αλγόριθμοι

Το μήκος (σε κατάλοιπα) αυτών των λέξεων καθορίζει και την ευαισθησία της αρχικής έρευνας: όσο μεγαλύτερο είναι το μήκος της λέξης, τόσο πιο απίθανο θα είναι να βρεθούν οι ίδιες λέξεις στις βάσεις κατά τύχη (δηλ. να είναι παρούσες σε μη ομόλογες αλληλουχίες). Ταυτόχρονα όμως αυξάνει η πιθανότητα μια αληθώς ομόλογη αλληλουχία να διαφύγει της έρευνας (ιδιαίτερα για αλληλουχίες με μακρινή εξελικτική σχέση). Για το λόγο αυτό το μήκος των λέξεων είναι συνήθως μικρό (π.χ. 2 κατάλοιπα για το FASTA).

Ευρεστικοί αλγόριθμοι

Επειδή αυτές οι 'λέξεις' είναι μικρού μήκους ακολουθίες, η αναζήτηση τους στις αλληλουχίες των βάσεων δεδομένων (χωρίς την εισαγωγή κενών) μπορεί να γίνει πολύ γρήγορα. Το αποτέλεσμα από μια τέτοια έρευνα (με βάση τις 'λέξεις') περιέχει θόρυβο, δηλαδή πολλές αλληλουχίες οι οποίες συμπτωματικά έχουν κάποια ή κάποιες κοινές λέξεις με την αλληλουχία-στόχος. Για το λόγο αυτό, τόσο το FASTA όσο και το BLAST χρησιμοποιούν επιπλέον κριτήρια για να περιορίσουν περαιτέρω τον αριθμό των αλληλουχιών που βρίσκουν από την πρώτη έρευνα.

BLAST

Το πρόγραμμα BLAST είναι το de facto standard για την έρευνα των βάσεων δεδομένων και θα παρουσιαστεί αναλυτικότερα. Το BLAST δεν απαιτεί οι λέξεις να είναι ταυτόσημες :
Αυτό που κάνει είναι να υπολογίζει για κάθε λέξη (από την αλληλουχία-στόχος) και για κάθε λέξη (από τις αλληλουχίες των βάσεων) μια βαθμολογία με βάση τους πίνακες βαθμολόγησης που έχουμε αναφέρει (χωρίς την εισαγωγή κενών). Εάν η βαθμολογία ξεπερνάει ένα προκαθορισμένο όριο, τότε η αντίστοιχη αλληλουχία (από τη βάση δεδομένων) περνάει στο επόμενο στάδιο του αλγόριθμου, αλλιώς απορρίπτεται.

BLAST

Όσες αλληλουχίες πέρασαν το πρώτο στάδιο επιλογής, εισέρχονται σε ένα δεύτερο στάδιο στο οποίο γίνεται μία απόπειρα να επεκταθούν οι στοιχίσεις των λέξεων ώστε να δημιουργηθεί μία τοπικά βέλτιστη στοίχιση μεταξύ των αλληλουχιών. Η επέκταση γίνεται και προς τις δυο κατευθύνσεις μέσω της χρήσης των πινάκων βαθμολόγησης και της εισαγωγής κενών. Εάν αυτή η επέκταση δεν καταφέρει να οδηγήσει σε μια τοπική στοίχιση με βαθμολογία μεγαλύτερη από ένα προκαθορισμένο όριο, τότε πάλι η αντίστοιχη αλληλουχία (από τη βάση δεδομένων) απορρίπτεται.

BLAST

Στο στάδιο της επέκτασης της στοίχισης γίνεται και ένας επιπλέον έλεγχος : εάν ο ρυθμός με τον οποίο μειώνεται η βαθμολογία της στοίχισης καθώς αυτή επεκτείνεται είναι τέτοιος ώστε να ώστε να γίνει απίθανο να φτάσει η στοίχιση το όριο βαθμολογίας του δεύτερου σταδίου, τότε πάλι η αλληλουχία απορρίπτεται. Αυτός ο μηχανισμός μειώνει το χρόνο που καταναλώνεται στην επέκταση των στοιχίσεων, αλλά μπορεί να οδηγήσει στην απώλεια μερικών αληθώς ομολόγων αλληλουχιών.

BLAST : υποπρογράμματα

- BLASTP : ερευνά πρωτεϊνικές βάσεις, με μία πρωτεϊνική αλληλουχία-στόχο.
- BLASTN : ερευνά νουκλεοτιδικές βάσεις, με μία νουκλεοτιδική αλληλουχία-στόχο.
- BLASTX : ερευνά πρωτεϊνικές βάσεις, με μία νουκλεοτιδική αλληλουχία-στόχο. Έλεγχος και για τα 6 πλαίσια ανάγνωσης. Χρήσιμο για νέες αλληλουχίες και ESTs.
- TBLASTN : ερευνά νουκλεοτιδικές βάσεις, με μία πρωτεϊνική αλληλουχία-στόχο. Έλεγχος και για τα 6 πλαίσια ανάγνωσης. Χρήσιμο για την εύρεση μη ταυτοποιημένων (σχολιασμένων) αλληλουχιών.

Προβλήματα έρευνας βάσεων

Υπολογιστικά προβλήματα

Το βασικότερο πρόβλημα φαίνεται να είναι η ύπαρξη επαναλαμβανόμενων μοτίβων στην αλληλουχία-στόχος, για παράδειγμα περιοχές (της αλληλουχίας) πλούσιες σε όξινα ή βασικά κατάλοιπα, πολυ-προλίνες, κοκ.
Για τα νουκλεϊκά οξέα, αντίστοιχο πρόβλημα ανακύπτει από τις διάφορες υπο-οικογένειες των αλληλουχιών Alu.
Εάν τα αποτελέσματα μιας έρευνας δίνουν πολλές θετικές αλληλουχίες (hits) οι οποίες φαίνονται να είναι μεταξύ τους λειτουργικά ασύνδετες, ίσως να είναι σκόπιμο να εξεταστεί η αλληλουχία-στόχος για περιοχές μικρής πολυπλοκότητας (LCR για Low Complexity Regions).

Προβλήματα έρευνας βάσεων

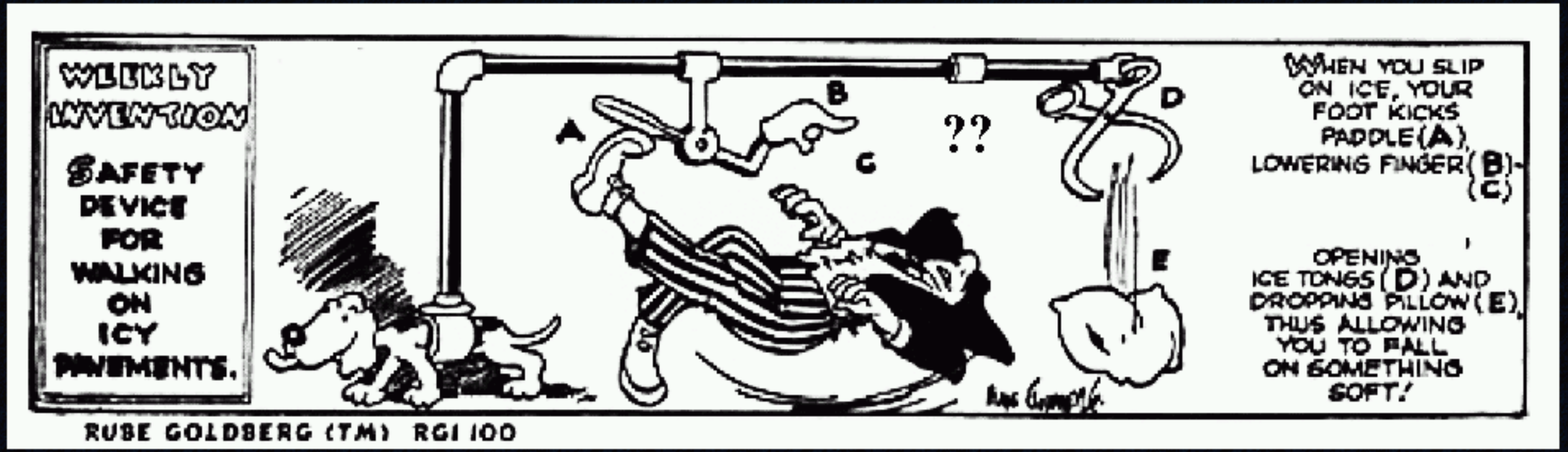
Υπολογιστικά προβλήματα

Μια προληπτική μέθοδος για την αποφυγή προβλημάτων με τα LCRs, είναι το φιλτράρισμα αυτών των περιοχών πριν την πραγματοποίηση της έρευνας.
Οι μέθοδοι που χρησιμοποιούνται για την εύρεση των LCRs ανήκουν στα ειδικά κεφάλαια. Εδώ αρκεί να σημειωθεί ότι τα δημοφιλέστερα προγράμματα για έρευνες βάσεων δεδομένων (BLAST) επιτρέπουν ως επιλογή (του χρήστη) το φιλτράρισμα τέτοιων περιοχών στην αλληλουχία-στόχος (όχι των αλληλουχιών της βάσης).

Προβλήματα έρευνας βάσεων

Βιολογικά προβλήματα

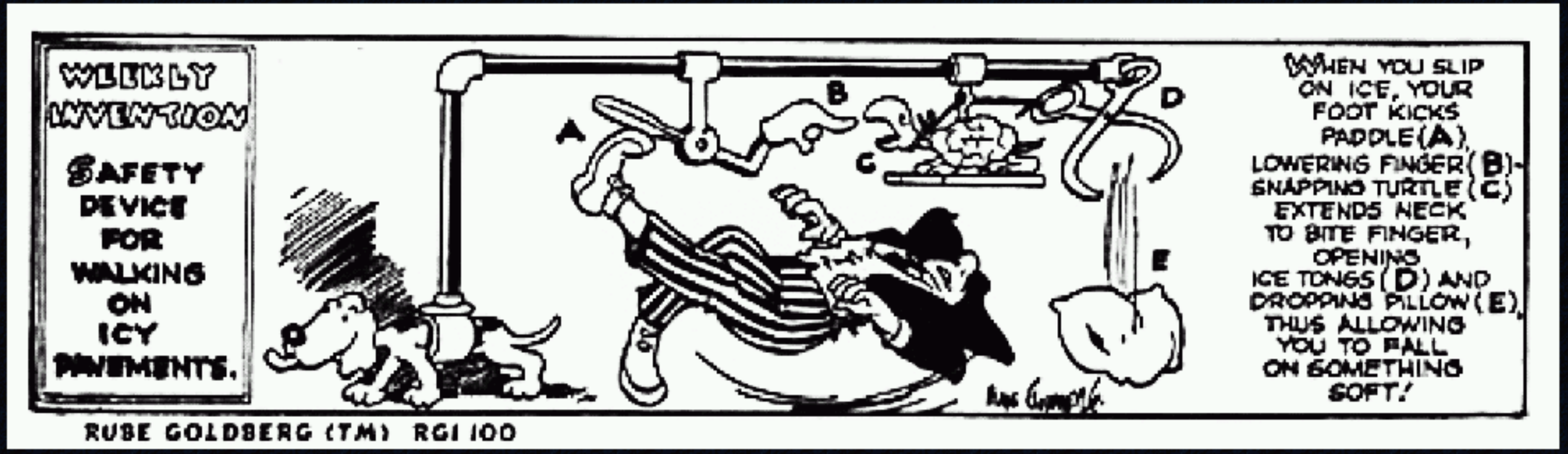
Το βασικότερο πρόβλημα είναι η γνώση των ορίων ερμηνείας των αποτελεσμάτων, ιδιαίτερα για πρωτεΐνες που αποτελούνται από πολλές υπομονάδες.



Προβλήματα έρευνας βάσεων

Βιολογικά προβλήματα

Το βασικότερο πρόβλημα είναι η γνώση των ορίων ερμηνείας των αποτελεσμάτων, ιδιαίτερα για πρωτεΐνες που αποτελούνται από πολλές υπομονάδες.



Στατιστική σημασία στοιχίσεων

Όπως προαναφέρθηκε, οι πίνακες βαθμολόγησης -με το να λαμβάνουν υπόψη την πιθανότητα τυχαίας στοίχισης δυο καταλοίπων- εμπεριέχουν ένα μέτρο της στατιστικής σημασίας των στοιχίσεων που προκύπτουν από τη χρήση τους. Με αυτό όμως τον τρόπο αγνοούμε την συχνότητα εμφάνισης των καταλοίπων στις συγκεκριμένες αλληλουχίες που εξετάζονται. Το ζητούμενο είναι να βρεθεί μια μέθοδος με βάση την οποία θα μπορούμε να υπολογίσουμε την στατιστική σημασία μίας συγκεκριμένης στοίχισης.

Στατιστική σημασία στοιχίσεων

Ιδανικά, θα θέλαμε δεδομένης μίας στοίχισης με βαθμολογία S να υπολογίσουμε την πιθανότητα E να επιτευχθεί τυχαία μια στοίχιση με την ίδια ή καλύτερη βαθμολογία. Εάν μία τέτοια μέθοδος υπήρχε και για κάποια στοίχιση είχαμε, για παράδειγμα, E=0.0010 τότε θα συνάγαμε ότι η πιθανότητα να επιτευχθεί κατά τύχη μια στοίχιση με την ίδια (ή καλύτερη) βαθμολογία είναι μόλις 1 στις 1000, και συνεπώς η στοίχιση είναι (μάλλον) στατιστικά σημαντική.

Στατιστική σημασία στοιχίσεων

Για να γίνει κάτι τέτοιο εφικτό, θα πρέπει να γνωρίζουμε την κατανομή βαθμολογιών που θα έδιναν οι τυχαίες στοιχίσεις. Αυτό έχει επιτευχθεί για την περίπτωση των τοπικών στοιχίσεων με τη χρήση της κατανομής ακραίων τιμών (extreme value distribution) η οποία θα αναπτυχθεί στα πλαίσια των ειδικών κεφαλαίων. Εδώ αρκεί να αναφέρουμε ότι είναι αυτές ακριβώς οι τιμές E που περιγράψαμε προηγουμένως οι οποίες δίδονται στην έξοδο που παράγεται τόσο από το πρόγραμμα BLAST όσο και από το FASTA.

Στατιστική ολικών στοιχίσεων

Η κατανομή ακραίων τιμών που αναφέραμε ισχύει μόνο για τοπικές στοιχίσεις. Για τις ολικές στοιχίσεις υπάρχει σαν εναλλακτική λύση η χρονοβόρα διαδικασία των Z-τιμών. Η διαδικασία είναι η εξής :

- Τα αμινοξέα της κάθε αλληλουχίας επαναδιευθετούνται με τυχαίο τρόπο (δηλαδή αλλάζουμε τη σειρά τους με τυχαίο τρόπο. Το μήκος και η σύσταση των αλληλουχιών παραμένουν αμετάβλητα).
- Οι νέες αλληλουχίες στοιχίζονται μεταξύ τους, και η προκύπτουσα βαθμολογία της στοίχισης σημειώνεται [ας την ονομάσουμε Ran(1)].

Στατιστική ολικών στοιχίσεων

- Η διαδικασία επαναλαμβάνεται κυκλικά [και έτσι παίρνουμε τα Ran(2), Ran(3), ..., Ran(N)].

Από τα Ran(1) ... Ran(N) υπολογίζουμε τη μέση τιμή (R) και την τυπική απόκλιση σ(R) της κατανομής βαθμολογιών των τυχαίων στοιχίσεων.

Η στατιστική σημαντικότητα της αρχικής στοίχισης υπολογίζεται ως Z-τιμή, Z = (S - R) / σ(R),
όπου S είναι η βαθμολογία της αρχικής στοίχισης.
Τιμές του Z μεγαλύτερες του 10 (?) θεωρούνται σημαντικές.

Εργαλεία δικτύου

SRS@ EMBL-EBI

Quick Search | Library Page | Query Form | Tools | Results | Projects | Views | Databanks | HELP

SRS

Quick Launch

Launch analysis tool:

antigenic

Launch

Packages Information

BLAST
FASTA
CLUSTAL
OTHER
EMBOSS

Available Analysis Tools - listed by type

Expand all Collapse all

Alignment Tools

- Alignment Consensus
- Alignment Differences
- Alignment Dot Plots
- Alignment Global
- Alignment Local
- Alignment Multiple

Display Tools
Edit Tools
Information Tools
Nucleic Tools
Protein Tools
Phylogeny Tools
Similarity Search Tools

BlastP BLASTP Protein-Protein Sequence Similarity Search - Launch *More Info...*

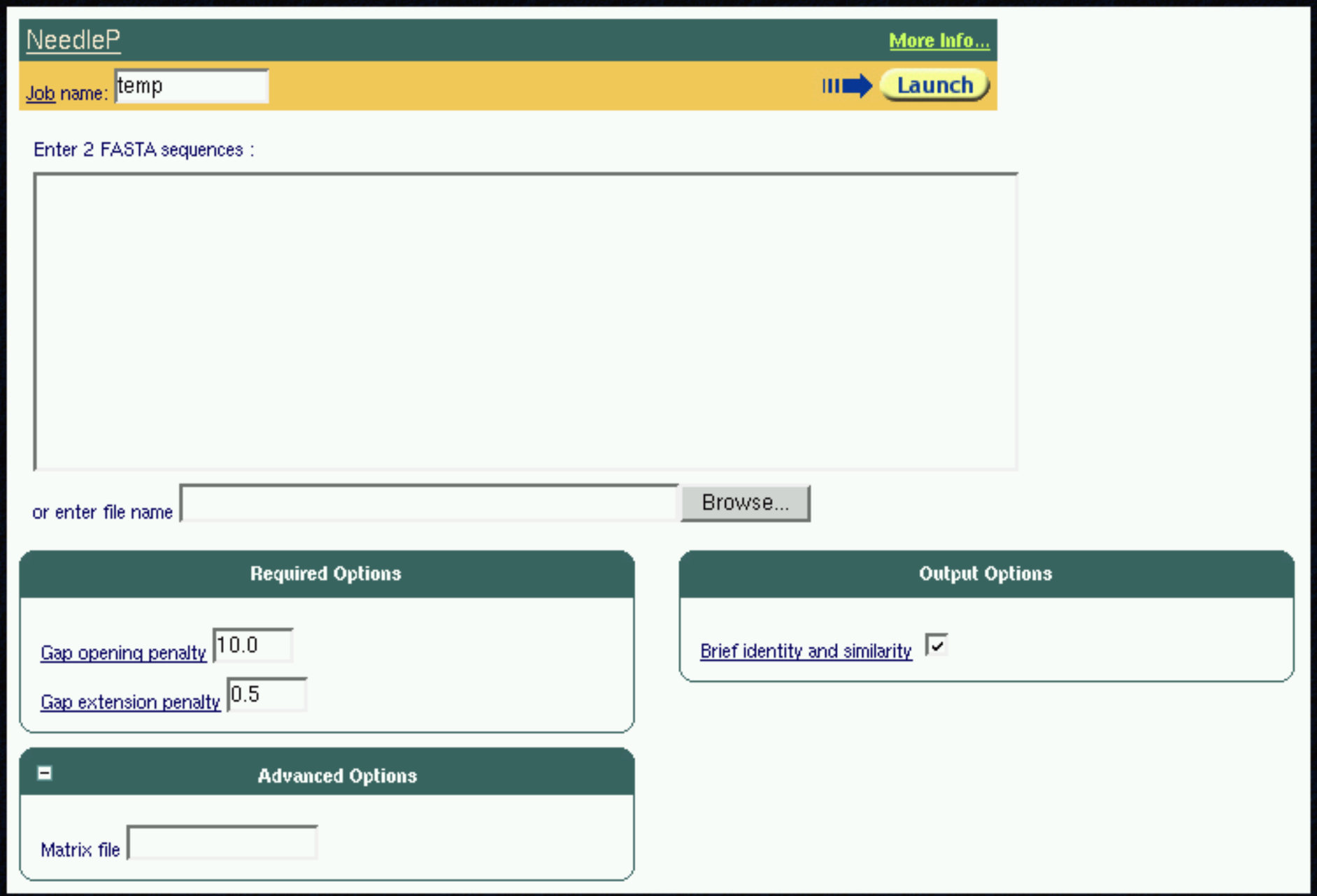
BlastN Sequence Similarity Search - Launch *More Info...*

FastA Sequence Similarity Search - Launch *More Info...*

FastX Sequence Similarity Search - Launch *More Info...*

FastY Sequence Similarity Search - Launch *More Info...*

Παραδείγματα : Needle



BLASTP : παράδειγμα χρήσης

Το πρόβλημα είναι να εντοπίσουμε πρωτεΐνες ομόλογες με κάποια πρωτεΐνη με το όνομα "BseCI". Χρησιμοποιούμε τις δυνατότητες που παρέχει το SRS για να βρούμε :

1. την αλληλουχία (και μια περιγραφή της λειτουργίας) αυτής της πρωτεΐνης, και,
2. το πρόγραμμα BLASTP για να βρούμε πιθανώς ομόλογες πρωτεΐνες της BseCI που να έχουν κατατεθεί στην SwissProt.

BLASTP : παράδειγμα χρήσης

BlastP — More Info...

Job name: temp Database to search: SWISSPROT (Release) Launch

SWALL (SPTR):MTC1_BACST

begin 1

end 579

```
1           11          21          31          41          51
MMSVQKANTVSRQKATGAHFTPDKLAEVIAKRILDYFKGEKNRVIRVLDPACGDGELLLA
61          71          81          91          101         111
INKVAQSMNIQLELIGVDFDIDAINIANERLSRSGHKNFRLINKDFLEMVSEGDNYDLFN
121         131         141         151         161         171
IEELEPVDIIIANPPYVRTQILGAEKAQKLREKFNLKGRVDLYQAFLVAMTQQLKSNGII
181         191         201         211         221         231
```

Output Options

Number of hits and alignments to show 50

Number of best hits from a region to keep 100

Search Parameters

Filter query sequence ☐

Scoring matrix BLOSUM62

The E value 10.000000

word size Default

Perform gapped alignment ☑

Cost to open a gap Default

Cost to extend a gap Default