Αναζήτηση μακρινών ομολόγων και στοίχιση profile-profile

Παντελης Μπάγκος

2019

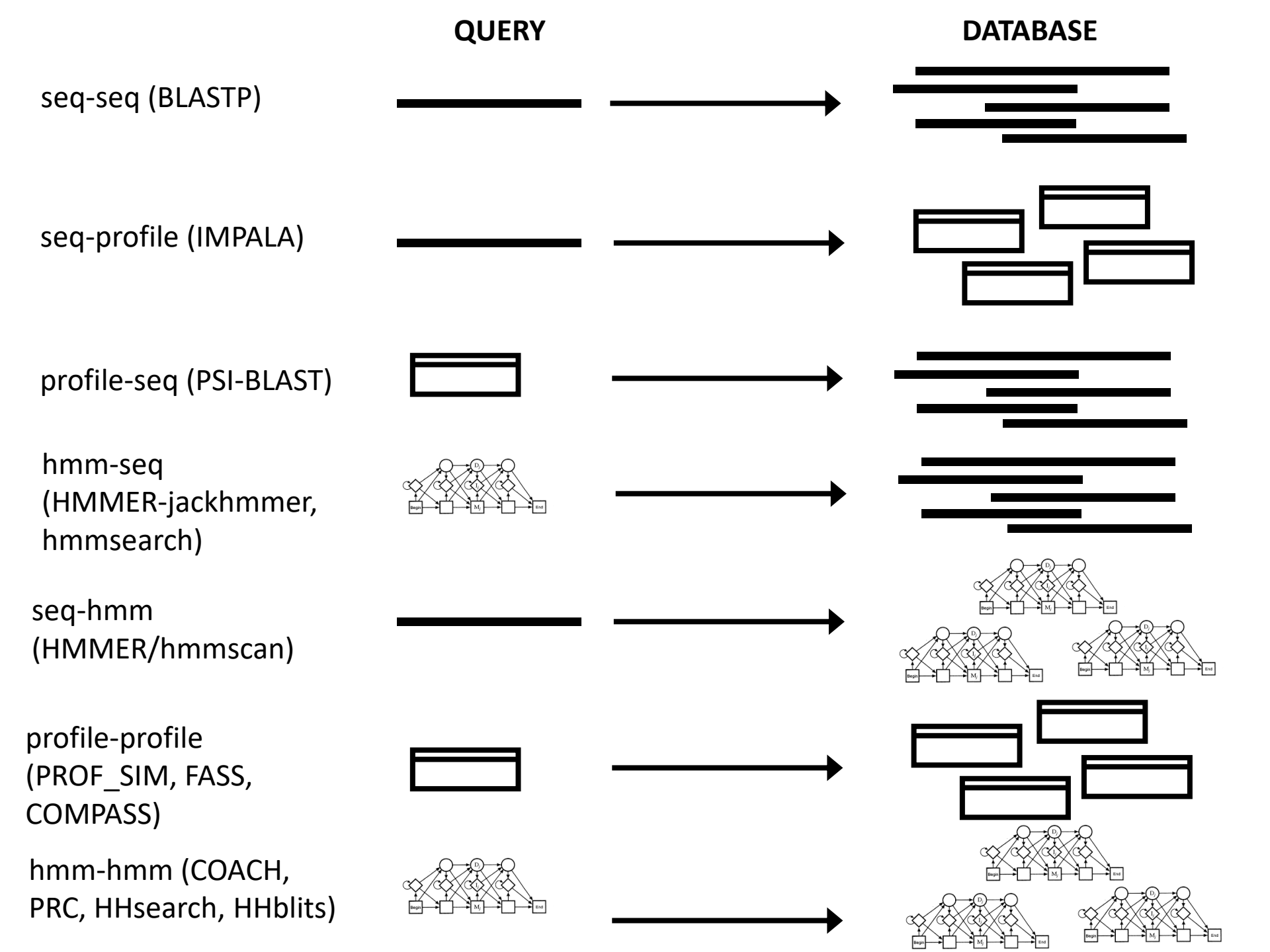
ΕΙΣΑΓΩΓΗ

Οι πλέον αποτελεσματικές μέθοδοι πρόβλεψης δομής και της λειτουργίας της δομής βασίζονται στην εύρεση σημαντικής ομολογίας/ομοιότητας μεταξύ της συγκεκριμένης πρωτεΐνης ενδιαφέροντος και μιας ήδη χαρακτηρισμένης πρωτεΐνης από μια βάση δεδομένων. Οι συνήθεις μέθοδοι σύγκρισης αλληλουχιών, όμως, χάνουν γρήγορα την ευαισθησία τους στην λεγόμενη «ζώνη του λυκόφωτος», δηλαδή σε περιπτώσεις πρωτεϊνών με 30% ή μικρότερη ομοιότητα σε επίπεδο αλληλουχίας.

Η ευαισθησία της αναγνώρισης των ομολόγων μπορεί να βελτιωθεί με τη χρήση πληροφοριών που εμπεριέχονται στις οικογένειες πρωτεϊνικών αλληλουχιών που συνδέονται με ανιχνεύσιμη ομολογία. Σε ένα πρώτο επίπεδο αυτής της προσέγγισης, συγκρίνεται μια αλληλουχία πρωτεΐνης με μια οικογένεια πρωτεϊνών που αντιπροσωπεύεται από ένα προφίλ [π.χ. στο PSI-BLAST ή στο HMMER]. Ένα επόμενο βήμα στη στρατηγική αυτή είναι η σύγκριση δύο προφίλ αλληλουχιών, είτε με τη μορφή PSSM, είτε με τη μορφή μοντέλων HMM, ενώ έχουν προταθεί και ευριστικές προσεγγιστικές λύσεις.

Κατηγορίες μεθόδων

Κλασικές μέθοδοι στοίχισης αλληλουχίας-profile	Περιλαμβάνονται οι επεκτάσεις του BLAST και οι υλοποιήσεις HMM για ευαίσθητες αναζητήσεις
Κλασικές μέθοδοι στοίχισης profile-profile	Στις μεθόδους αυτές περιλαμβάνονται οι μεθοδολογίες που στοιχίζουν μεταξύ τους προφίλ αλληλουχιών ή PSSM
Μέθοδοι στοίχισης HMM-HMM	Η κατηγορία περιλαμβάνει τις πιο εξελιγμένες μεθόδους που χρησιμοποιούν ειδικούς αλγόριθμους για να στοιχίσουν ολόκληρα HMM
Προσεγγιστικές ή ευριστικές μέθοδοι	Μέθοδοι που προσεγγίζουν το πρόβλημα, συνήθως μετατρέποντας ξανά το προφίλ σε μια «ψευτο»ακολουθία


QUERY
DATABASE
seq-seq (BLASTP)
seq-profile (IMPALA)
profile-seq (PSI-BLAST)
hmm-seq
(HMMER-jackhmmer,
hmmsearch)
seq-hmm
(HMMER/hmmscan)
profile-profile
(PROF_SIM, FASS,
COMPASS)
hmm-hmm (COACH,
PRC, HHsearch, HHblits)

Κλασικές μέθοδοι στοίχισης αλληλουχίας-profile

Περιλαμβάνονται οι επεκτάσεις του BLAST και οι υλοποιήσεις HMM για ευαίσθητες αναζητήσεις

- PSI-PLAST (Position-specific-iterated BLAST)
- PHI-BLAST (pattern-hit initiated BLAST)
- CS-BLAST (context-specific BLAST)
- DELTA-BLAST (domain enhanced lookup time accelerated BLAST)
- IMPALA (Integrating Matrix Profiles and Local Alignments)
- HMMER
- SAM

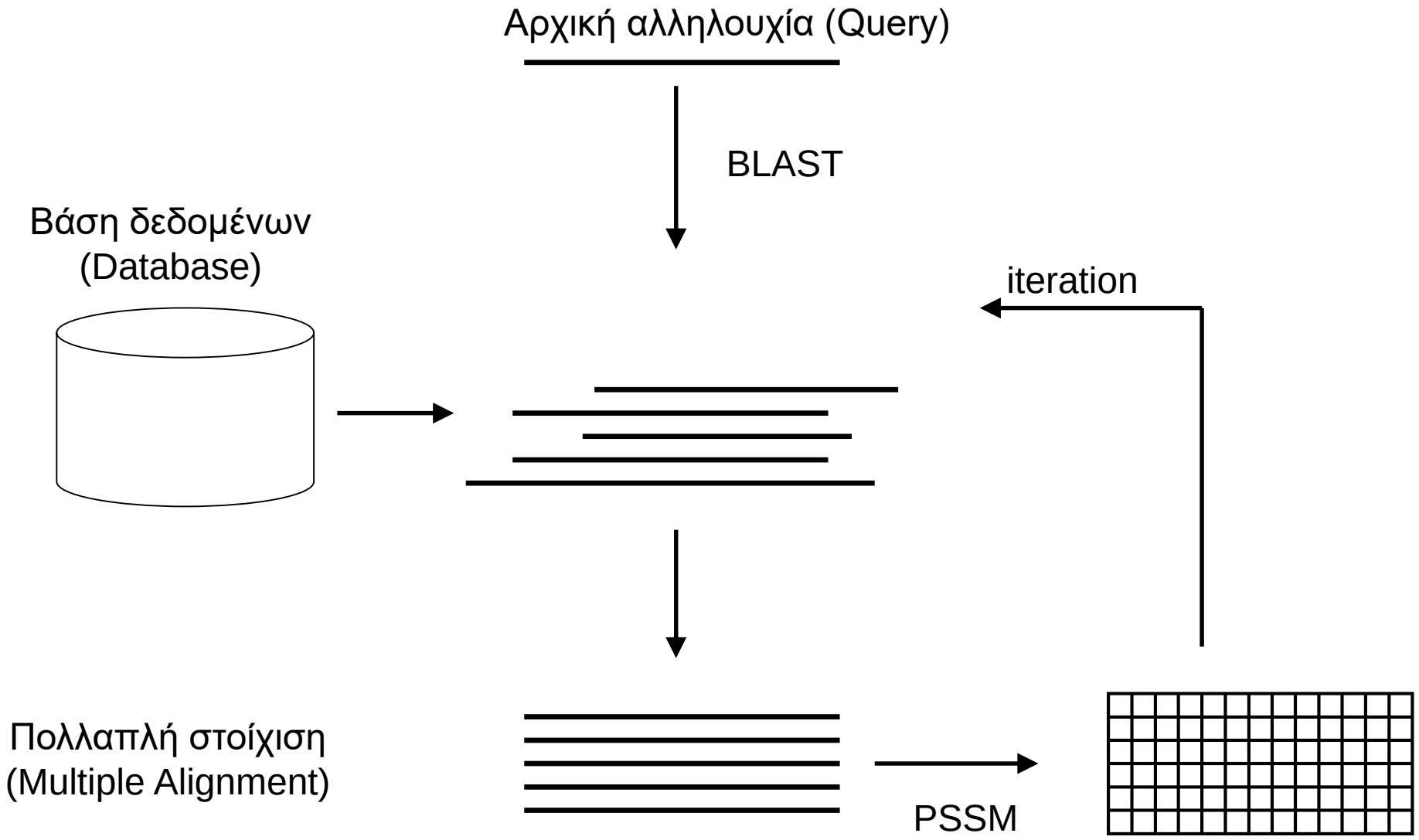
PSI-BLAST (Position-specific-iterated BLAST)

Είναι μια επέκταση του γνωστού αλγορίθμου BLAST και χρησιμοποιείται για την εύρεση μακρινών ομολόγων. Η μέθοδος δουλεύει ως εξής:

- Στην αρχή πραγματοποιείται μια κανονική αναζήτηση με το BLAST και συλλέγονται οι αλληλουχίες με E-value μικρότερο από κάποιο όριο που ορίζεται από τον χρήστη.
- Αυτές θεωρείται ότι είναι οι «σίγουρες» ομόλογες και χρησιμοποιούνται για να κατασκευαστεί ένας PSSM όπως περιγράψαμε παραπάνω, χωρίς όμως κενά καθώς κάθε στήλη του αντιστοιχεί σε μια θέση της αλληλουχίας της αρχικής πρωτεΐνης.
- Με αυτόν τον πίνακα, πραγματοποιείται εκ νέου αναζήτηση στη βάση δεδομένων, η οποία πλέον θα δώσει περισσότερες ομόλογες με E-value μικρότερο από το αρχικό όριο.

Η διαδικασία αυτή επαναλαμβάνεται αρκετές φορές, είτε μέχρι να σταματήσουν να προστίθενται νέες αλληλουχίες, είτε μέχρι να ξεπεραστεί ένας συγκεκριμένος αριθμός επαναλήψεων (συνήθως 3 ή 4).

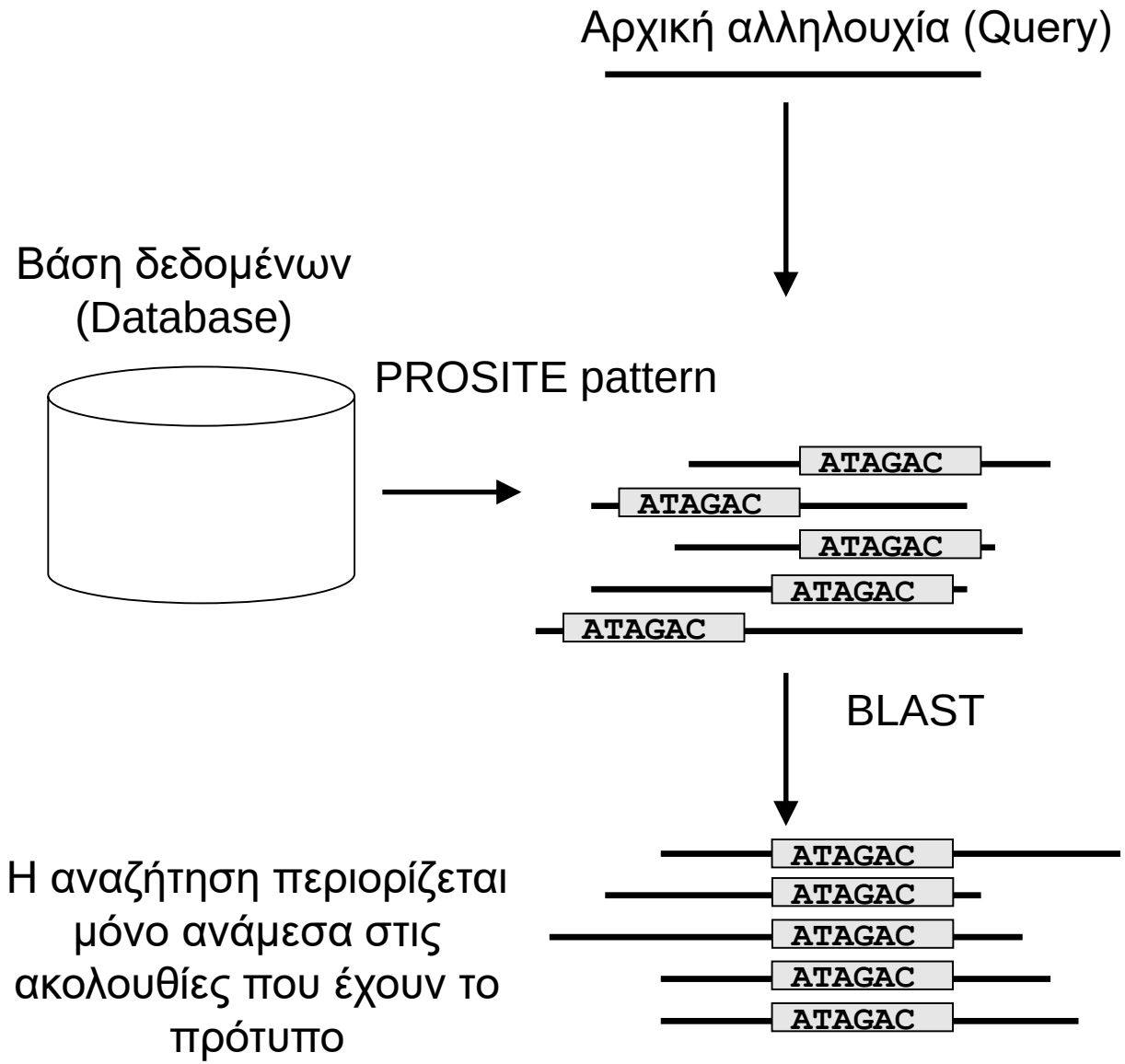
Η μέθοδος είναι εξαιρετικά αποδοτική και εντοπίζει μεγάλο αριθμό ομόλογων πρωτεϊνών (μακρινών ομολόγων), οι οποίες δεν θα μπορούσαν να εντοπιστούν με μια συμβατική αναζήτηση. Η επαναληπτική αυτή διαδικασία, θυμίζει τον αλγόριθμο EM, και οι μόνες περιπτώσεις στις οποίες μπορεί να αποτύχει είναι είτε όταν δεν βρεθούν καθόλου ομόλογες στην πρώτη αναζήτηση, είτε όταν το όριο είναι αρκετά ψηλά με συνέπεια να συμπεριληφθούν και πρωτεΐνες που δεν έχουν πραγματική ομολογία, οπότε και το προφίλ δεν θα είναι πλέον ειδικό αρκετά (contamination).


Αρχική αλληλουχία (Query)
BLAST
Βάση δεδομένων
(Database)
iteration
Πολλαπλή στοίχιση
(Multiple Alignment)
PSSM

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

PHI-BLAST (pattern-hit initiated BLAST)

Το **PHI-BLAST** (pattern-hit initiated BLAST) είναι άλλη μια παραλλαγή του BLAST, η οποία όμως χρησιμοποιεί πρότυπα κανονικών εκφράσεων (Zhang et al., 1998). Η ιδέα εδώ είναι διαφορετική και συνίσταται στη χρησιμοποίηση γνωστών πρότυπων, τα οποία υπάρχουν στην αλληλουχία επερώτησης και τα καθορίζει ο χρήστης, για να καθοδηγήσουν την αναζήτηση. Με τον τρόπο αυτό, το εύρος της αναζήτησης περιορίζεται και σε πολλές περιπτώσεις εντοπίζονται ομόλογες πρωτεΐνες οι οποίες δεν μπορούσαν να εντοπιστούν με το συμβατικό τρόπο αναζήτησης


Αρχική αλληλουχία (Query)
Βάση δεδομένων (Database)
PROSITE pattern
ATAGAC
ATAGAC
ATAGAC
ATAGAC
ATAGAC
BLAST
ATAGAC
ATAGAC
ATAGAC
ATAGAC
ATAGAC
Η αναζήτηση περιορίζεται μόνο ανάμεσα στις ακολουθίες που έχουν το πρότυπο

CS-BLAST

Οι τυποποιημένες μέθοδοι σύγκρισης αλληλουχιών χρησιμοποιούν πίνακες υποκατάστασης για να βρουν την στοίχιση με το καλύτερο άθροισμα βαθμολογιών ομοιότητας μεταξύ στοιχισμένων καταλοίπων. Αυτά τα σκορ ομοιότητας δεν λαμβάνουν υπόψη το τοπικό πλαίσιο της αλληλουχίας. Η μέθοδος αυτή προτείνει μια προσέγγιση που εξάγει ομοιότητες αμινοξέων από κοντινά παράθυρα επικεντρωμένα σε κάθε κατάλοιπο της αλληλουχίας επερώτησης. Τα αποτελέσματά καταδεικνύουν ότι το πλαίσιο της αλληλουχίας, τα γειτονικά κατάλοιπα δηλαδή, περιέχει πολύ περισσότερες πληροφορίες σχετικά με τις αναμενόμενες μεταλλάξεις απ 'ό, τι το ίδιο το αμινοξικό κατάλοιπο. Χρησιμοποιώντας τις ομοιότητες που σχετίζονται με το περιβάλλον (CS-BLAST) σε συνδυασμό με το κλασικό NCBI BLAST, αυξάνεται η ευαισθησία περισσότερο από 2 φορές σε ένα δύσκολο σετ αναφοράς, χωρίς απώλεια της ταχύτητας. Η ποιότητα της στοίχισης βελτιώνεται επίσης σημαντικά. Επιπλέον, σημαντικές βελτιώσεις επιτυγχάνονται κατά την εφαρμογή αυτού του προτύπου στα προφίλ αλληλουχίας: Δύο επαναλήψεις του CSI-BLAST, της έκδοσης PSI-BLAST που λαμβάνει υπόψη το περιβάλλον, είναι πιο ευαίσθητες από 5 επαναλήψεις του PSI-BLAST.

Biegert A, Söding J.Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A. 2009 Mar 10;106(10):3770-5. doi: 10.1073/pnas.0810767106. Epub 2009 Feb 20.

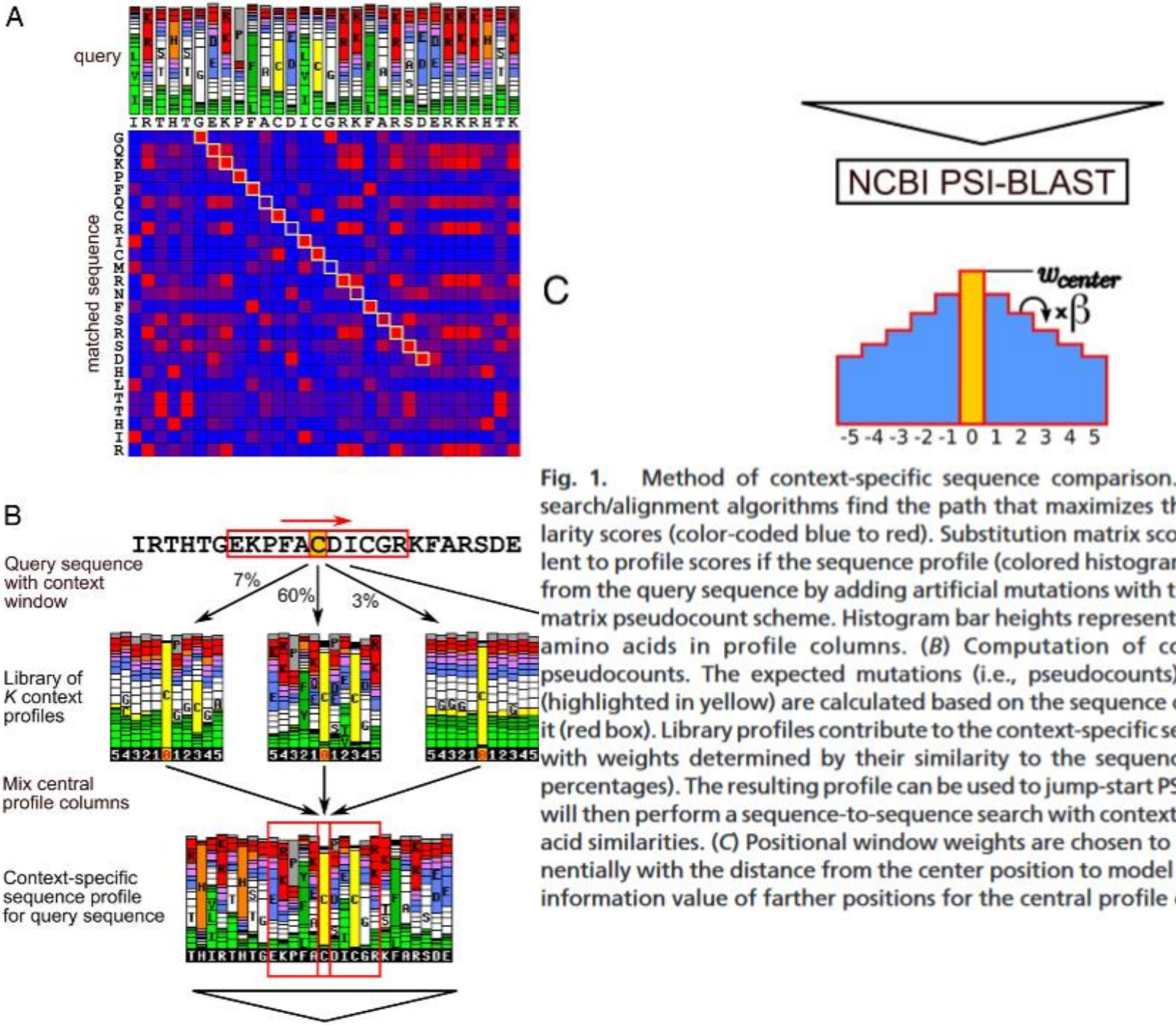


Fig. 1. Method of context-specific sequence comparison. (*A*) Sequence search/alignment algorithms find the path that maximizes the sum of similarity scores (color-coded blue to red). Substitution matrix scores are equivalent to profile scores if the sequence profile (colored histogram) is generated from the query sequence by adding artificial mutations with the substitution matrix pseudocount scheme. Histogram bar heights represent the fraction of amino acids in profile columns. (*B*) Computation of context-specific pseudocounts. The expected mutations (i.e., pseudocounts) for a residue (highlighted in yellow) are calculated based on the sequence context around it (red box). Library profiles contribute to the context-specific sequence profile with weights determined by their similarity to the sequence context (see percentages). The resulting profile can be used to jump-start PSI-BLAST, which will then perform a sequence-to-sequence search with context-specific amino acid similarities. (*C*) Positional window weights are chosen to decrease exponentially with the distance from the center position to model the decreasing information value of farther positions for the central profile column.

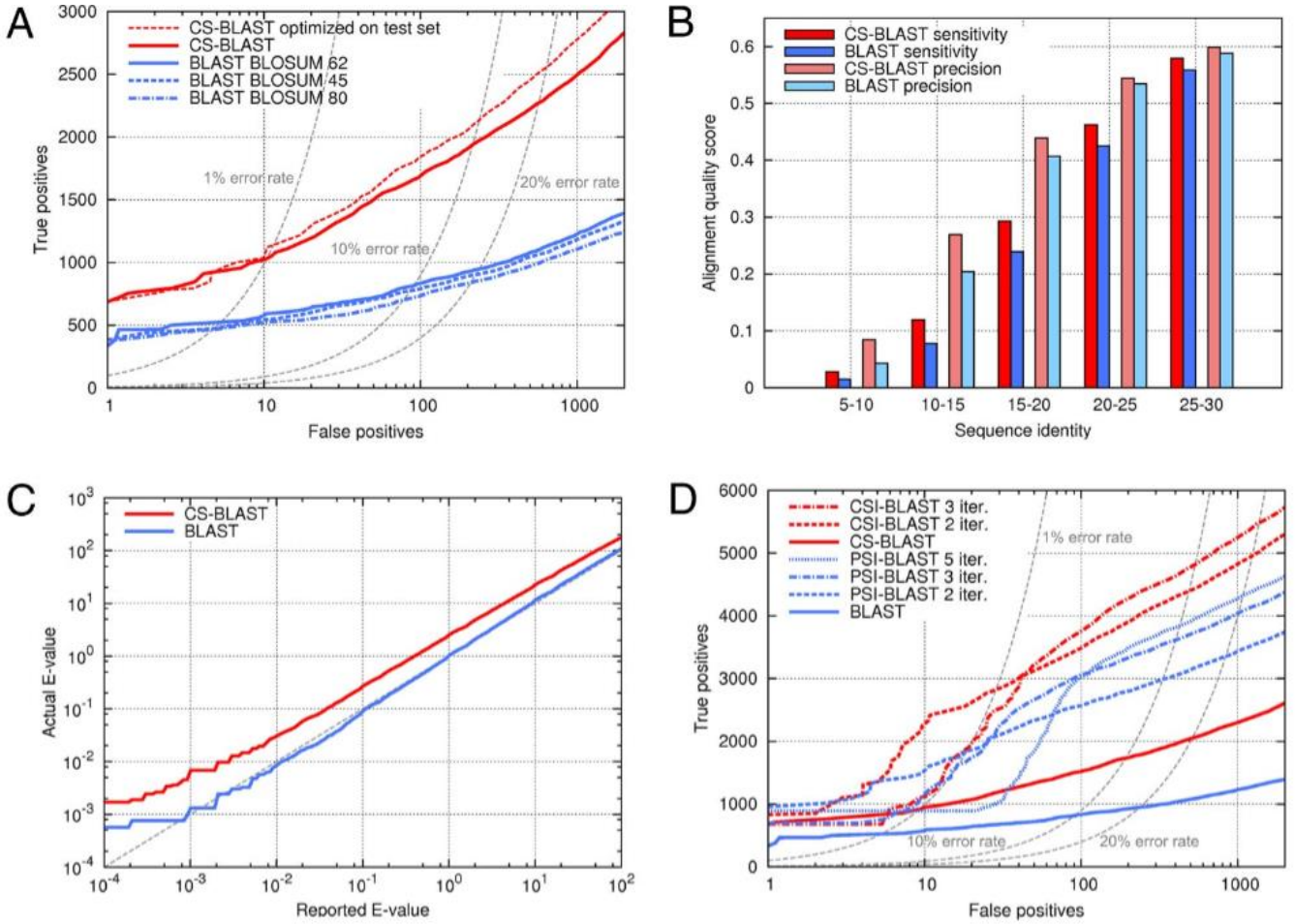


Fig. 2. Context information improves search performance and alignment quality. (*A*) Homology detection benchmark on SCOP20 dataset: true positives (pairs from the same SCOP superfamily) versus false positives (pairs from different folds). CS-BLAST detects 138% more true positives than BLAST at 10% error rate. (*B*) CS-BLAST has better average alignment sensitivity and precision than BLAST over the entire range of sequence identities of the aligned pairs. (*C*) Actual versus reported E-values on the SCOP20 dataset show that CS-BLAST E-values are too optimistic by a factor of 3 to 5. (*D*) Same benchmark as *A* (note different *y*-scales), but comparing CSI-BLAST with PSI-BLAST for one to five iterations. Two CSI-BLAST iterations are more sensitive than five PSI-BLAST iterations.

DELTA-BLAST

Το DELTA-BLAST (domain enhanced lookup time accelerated BLAST), έρχεται να δώσει μια λίγο διαφορετική προσέγγιση σε σχέση με το κλασικό PSI-BLAST. Η μέθοδος πραγματοποιεί αναζήτηση έναντι προκατασκευασμένων PSSMs πριν πραγματοποιήσει την κανονική αναζήτηση στη βάση δεδομένων αλληλουχιών. Για τα PSSMs, το DELTA-BLAST χρησιμοποιεί ένα υποσύνολο του NCBI's Conserved Domain Database (CDD). Σε κάποια τεστ με δεδομένα της βάσης ASTRAL, με ένα γύρο αναζήτησης, το DELTA-BLAST πετυχαίνει καλύτερα αποτελέσματα ακόμα και σε σχέση με το CS-BLAST. Όταν πραγματοποιούνται περισσότερες επαναλήψεις το πλεονέκτημα μειώνεται αλλα σε κάθε περίπτωση το DELTA-BLAST εξακολουθεί να δίνει καλύτερα ROC scores από το CS-BLAST.

Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. Biol Direct. 2012 Apr 17;7:12. doi: 10.1186/1745-6150-7-12.

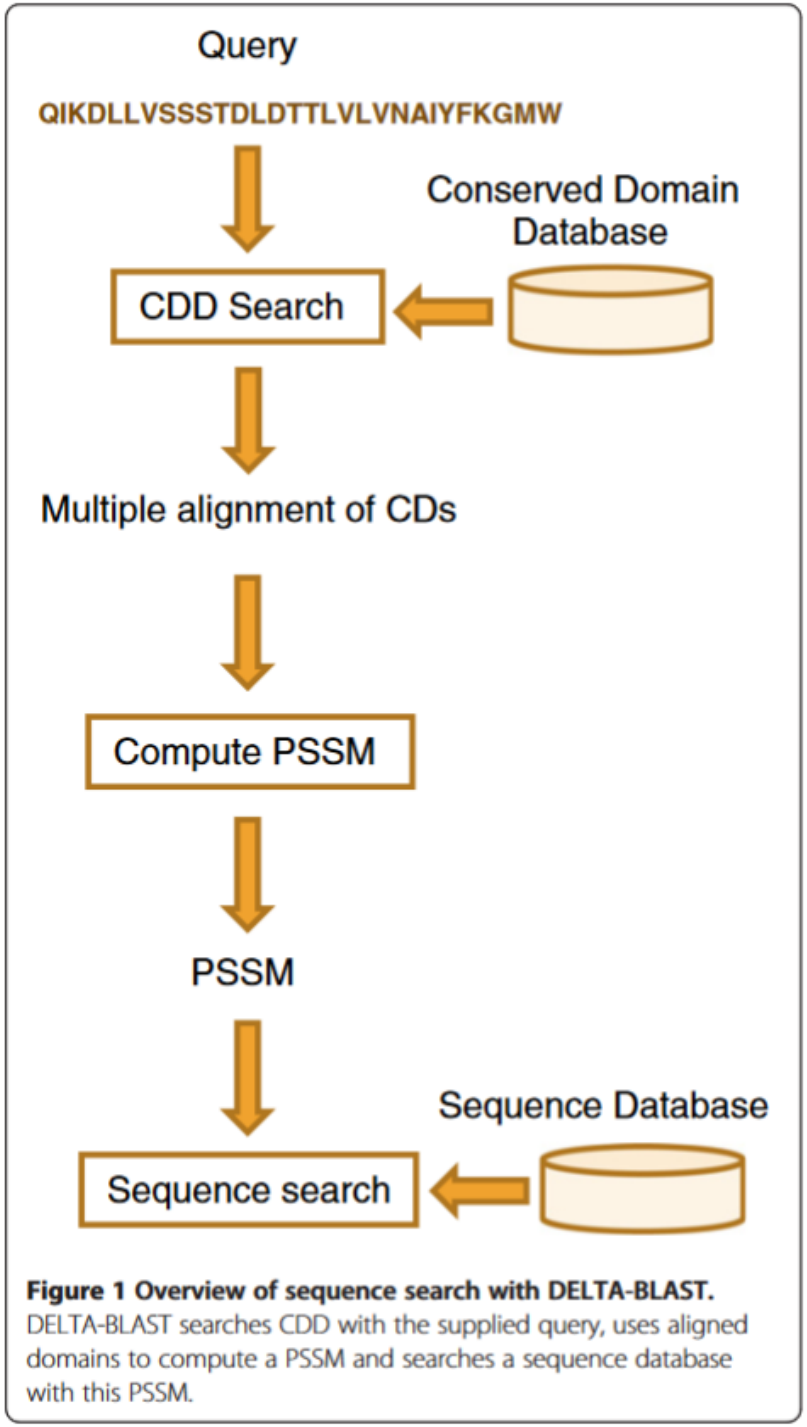


Figure 1 Overview of sequence search with DELTA-BLAST. DELTA-BLAST searches CDD with the supplied query, uses aligned domains to compute a PSSM and searches a sequence database with this PSSM.

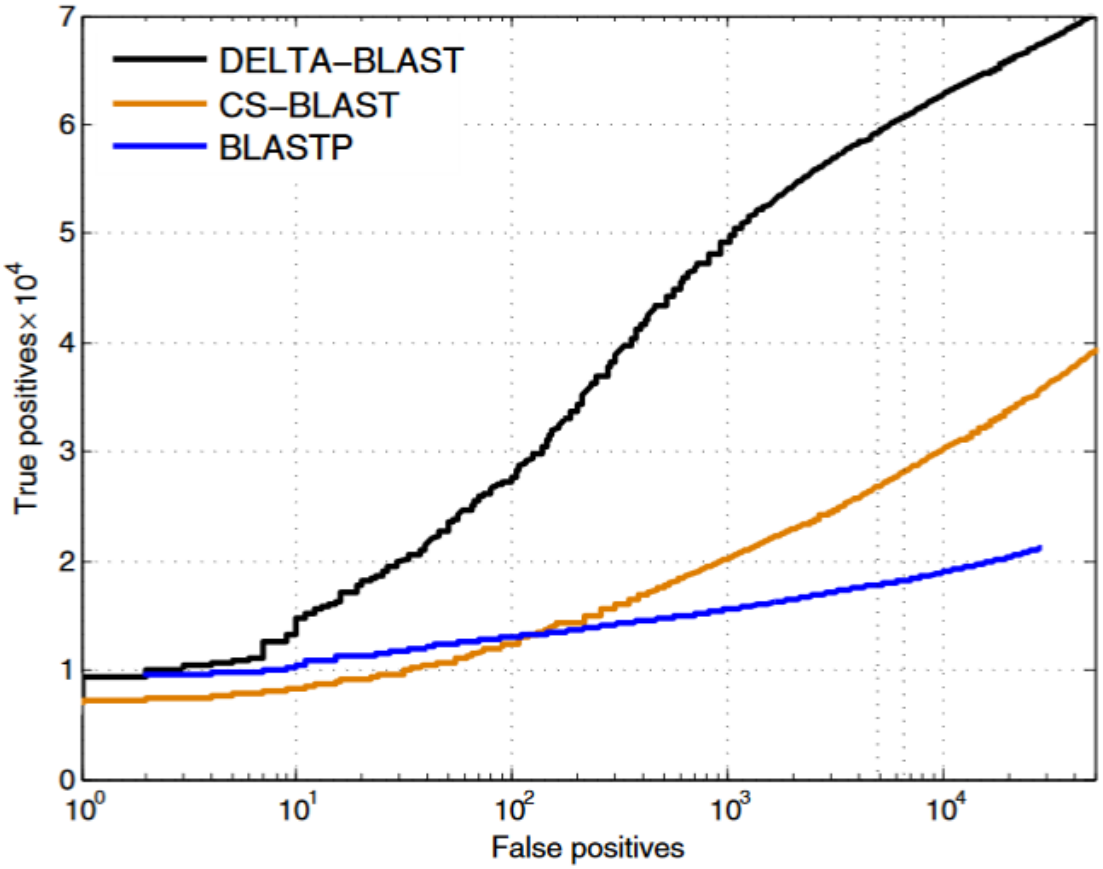


Figure 2 Number of true positives vs. number of false positives for DELTA-BLAST, CS-BLAST and BLASTP. The searched database was created using ASTRAL 40 sequences for SCOP version 1.75. To create the query set, we sorted the SCOP domains in lexicographic order and selected even numbered sequences for the test query set. We excluded from the query set any sequence that was the sole member of its superfamily in ASTRAL 40. We considered a query and database sequence to be homologs if they belonged to the same superfamily, and non-homologs if they belonged to different folds. The search results generated by all queries were pooled and ordered by E-value. The database and the query set consisted of 10,569 and 4852 sequences, respectively.

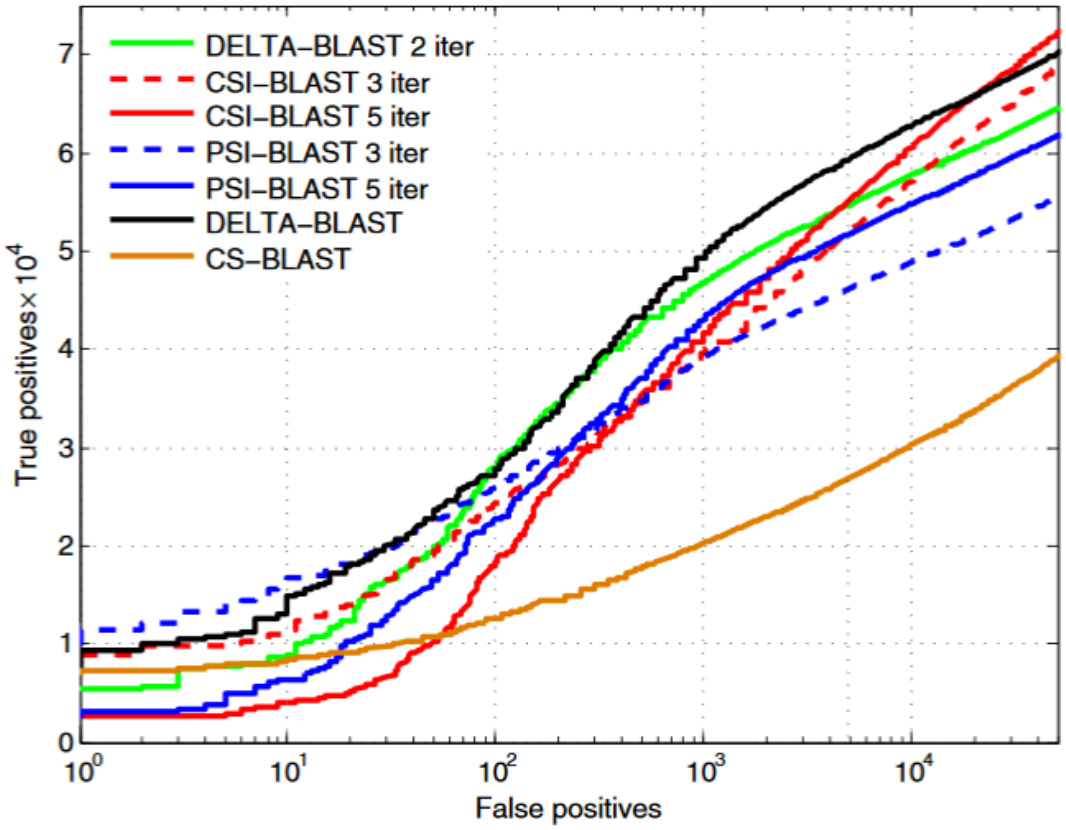


Figure 3 Number of true positives vs. number of false positives for PSI-BLAST, iterated DELTA-BLAST, CSI-BLAST, DELTA-BLAST, and CS-BLAST. See the legend of Figure 2.

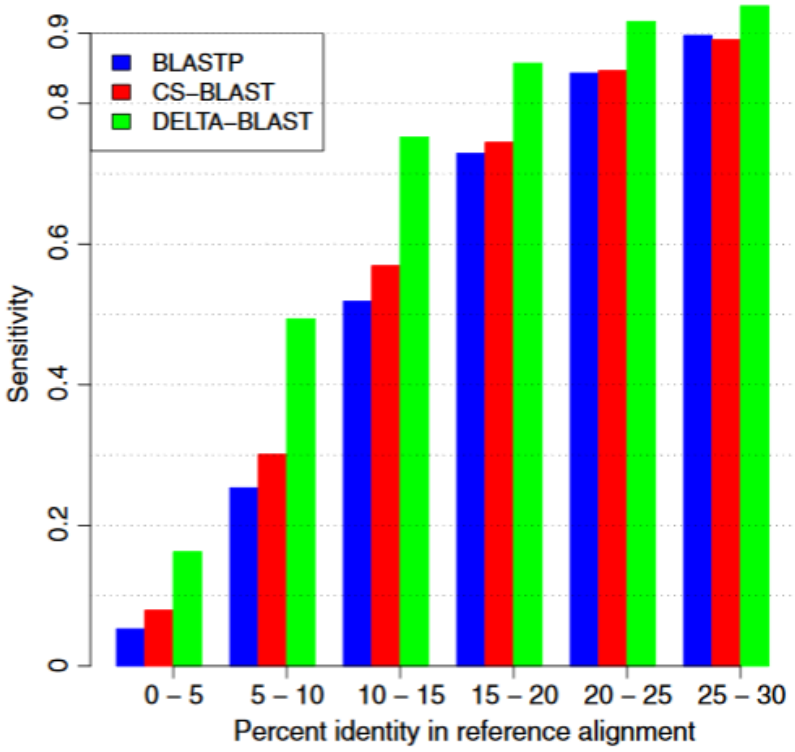


Figure 5 Alignment sensitivity of BLASTP, CS-BLAST, and DELTA-BLAST. Sensitivity measures the fraction of a reference alignment correctly recovered by a sequence alignment. Sequences and their reference alignments from the SABmark superfamily set were used to measure sensitivity. We used only reference alignments with sequence identity below 30% between sequences that did not correspond to SCOP domains present in the training set used to tune DELTA-BLAST parameters. Additionally, we removed reference alignments with fewer than five aligned pairs of residues. The data set contained 10,006 alignments between 2,379 sequences.

BLAST yields only a slightly larger percentage of TPs in this group than CS-BLAST. In general, Table 4 demonstrates common biases among all the search methods towards TPs represented in CDD.

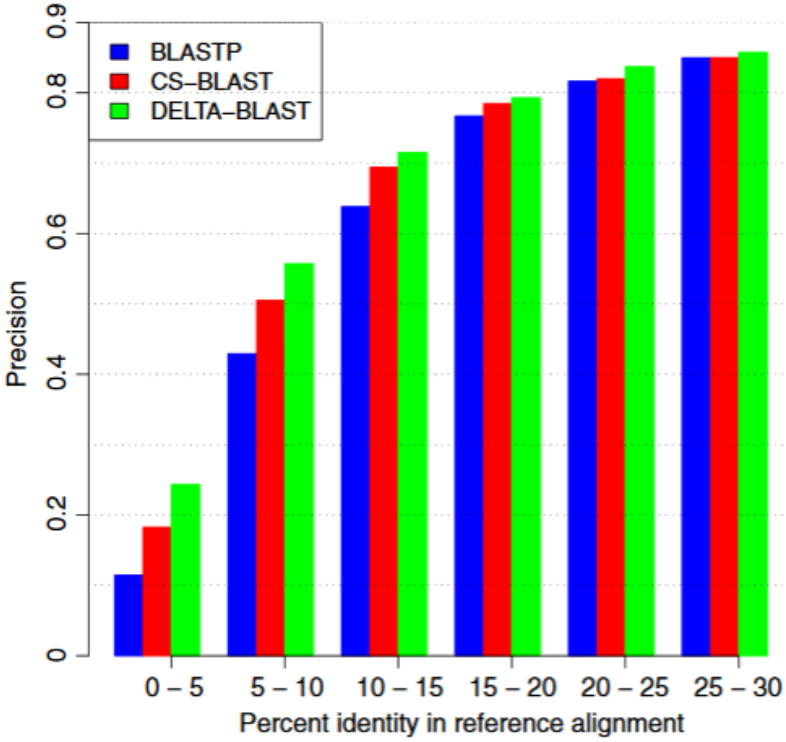
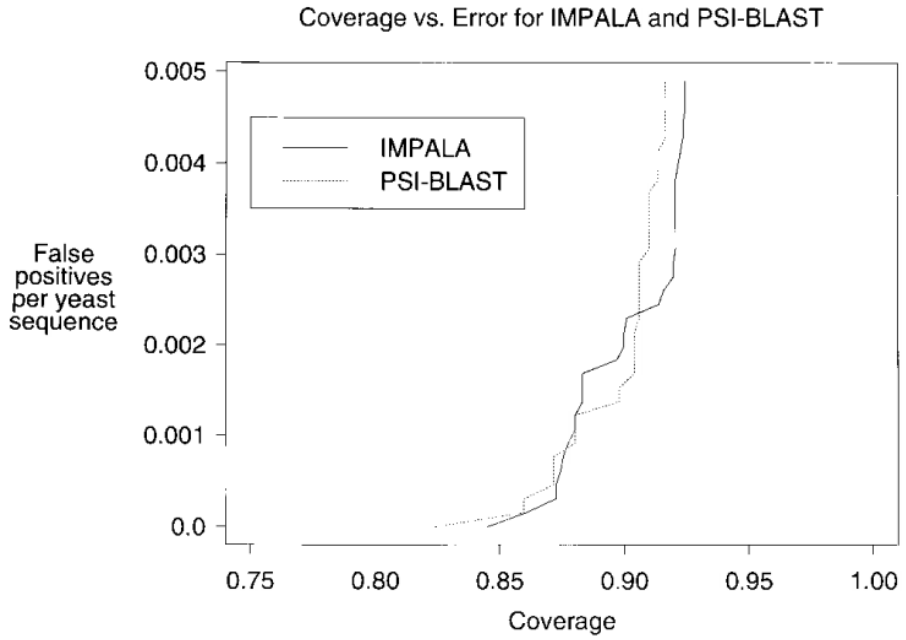


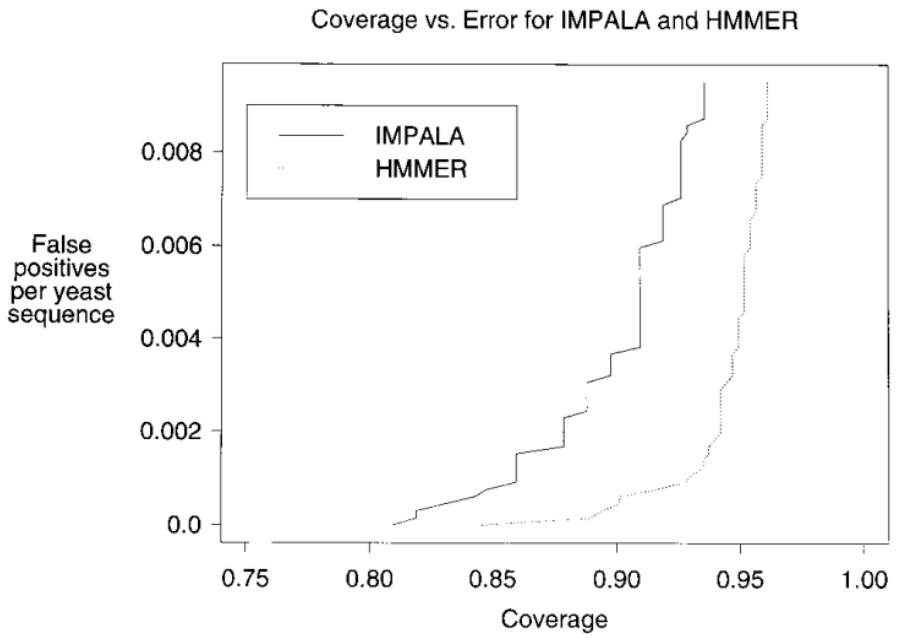
Figure 6 Alignment precision of BLASTP, CS-BLAST, and DELTA-BLAST. Precision measures the fraction of a sequence alignment that correctly reproduces a reference alignment. See the legend of Figure 5 for the data set description.

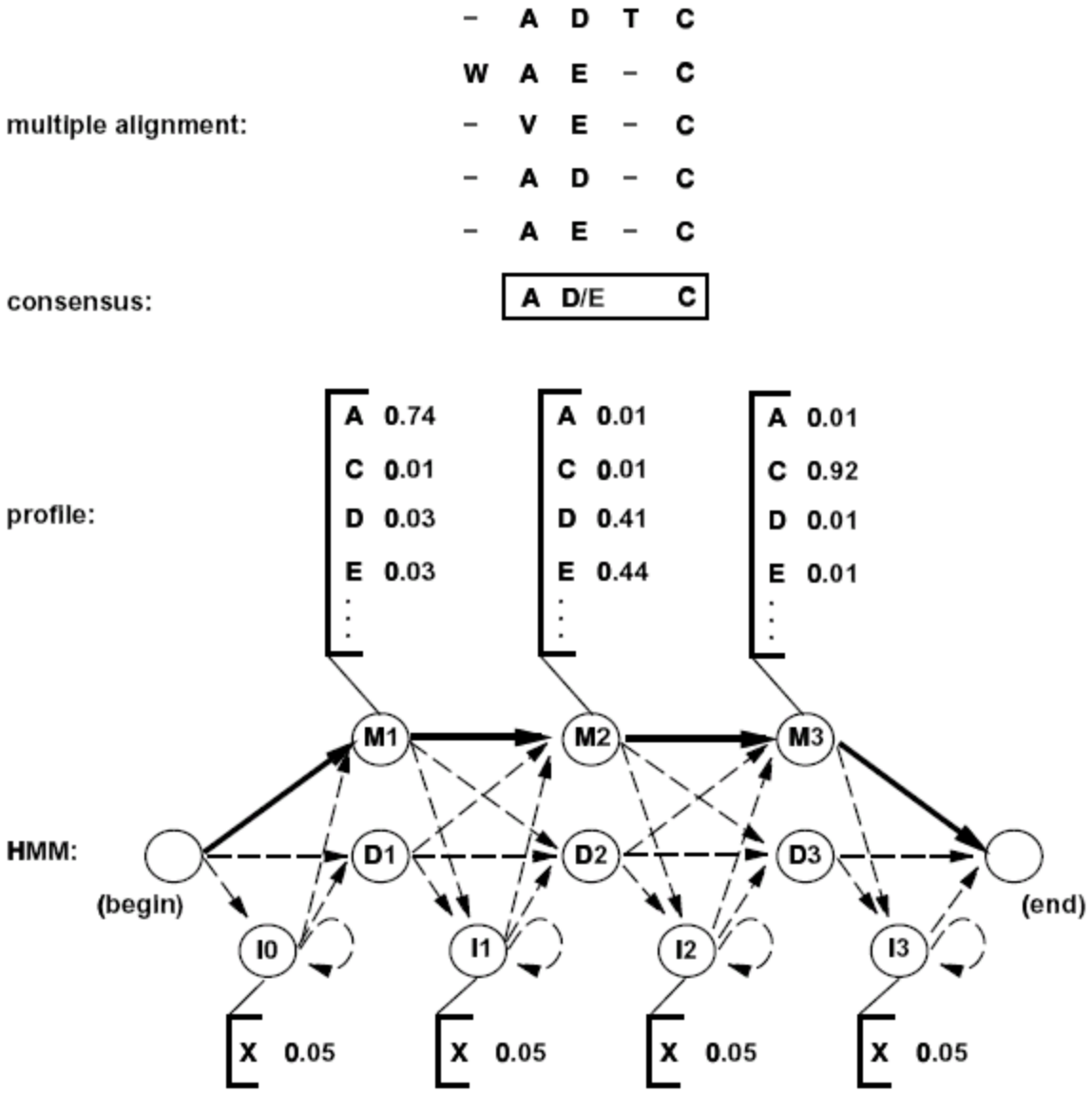
IMPALA

Σε αντίθεση με το PSI-BLAST το οποίο πραγματοποιεί αναζητήσεις ενός PSSM έναντι μιας βάσης δεδομένων αλληλουχιών, το IMPALA είναι ένα λογισμικό σχεδιασμένο να εκτελεί τη συμπληρωματική διαδικασία σύγκρισης, δηλαδή την αναζήτηση μίας μόνο αλληλουχίας επερώτησης απέναντι σε μια βάση δεδομένων PSSM που παράγεται από το PSI-BLAST. Η ευαισθησία του IMPALA σε αναζήτηση πρωτεϊνών με μακρινές βιολογικές ομοιότητες είναι σε γενικές γραμμές παρόμοια με αυτή του PSI-BLAST. Ωστόσο, το IMPALA χρησιμοποιεί καλύτερο τρόπο υπολογισμού της στατιστικής σημαντικότητας και, σε αντίθεση με το PSI-BLAST, εγγυάται την εύρεση της βέλτιστης τοπικής στοίχισης χρησιμοποιώντας τον αυστηρό αλγόριθμο Smith-Waterman. Επίσης, είναι αρκετά ταχύτερο σε μια μεγάλη βάση δεδομένων PSSM σε σχέση με το BLAST ή το PSI-BLAST όταν κάνει αναζήτηση στη βάση δεδομένων NR

Schäffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics. 1999 Dec;15(12):1000-11.


Coverage vs. Error for IMPALA and PSI-BLAST
False positives per yeast sequence
0.005
0.004
0.003
0.002
0.001
0.0
IMPALA
PSI-BLAST
0.75
0.80
0.85
0.90
0.95
1.00
Coverage


Coverage vs. Error for IMPALA and HMMER
False positives per yeast sequence
0.008
0.006
0.004
0.002
0.0
IMPALA
HMMER
0.75
0.80
0.85
0.90
0.95
1.00
Coverage


multiple alignment:
– A D T C
W A E – C
– V E – C
– A D – C
– A E – C
consensus:
A D/E C
profile:
A 0.74
C 0.01
D 0.03
E 0.03
A 0.01
C 0.01
D 0.41
E 0.44
A 0.01
C 0.92
D 0.01
E 0.01
HMM:
(begin)
M1
M2
M3
D1
D2
D3
(end)
I0
I1
I2
I3
X 0.05
X 0.05
X 0.05
X 0.05

Sequence profiles are a condensed representation of multiple alignments

master sequence →

```
HBA_human  ... W  G  K  V  G  A  -  -  H  A  G  E ...
HBB_human  ... W  G  K  V  -  -  -  -  N  V  D  E ...
MYG_phyca  ... W  G  K  V  E  A  -  -  D  V  A  G ...
LGB2_luplu ... W  K  D  F  N  A  -  -  N  I  P  K ...
GLB1_glydi ... W  E  E  I  A  G  A  D  N  G  A  G ...
```

Each column of the profile $p_j(a)$ contains the amino acid frequencies in the multiple sequence alignment

A	...	0	0	0	0	0.25	0.75		0	0.2	0.4	0	...
C	...	0	0	0	0	0	0		0	0	0	0	...
D	...	0	0	0.2	0	0	0		0.2	0	0.2	0	...
E	...	0	0.2	0.2	0	0.25	0		0	0	0	0.4	...
F	...	0	0	0	0.2	0	0		0	0	0	0	...
G	...	0	0.6	0	0	0.25	0.25		0	0.2	0.2	0.4	...
H	...	0	0	0	0	0	0		0.2	0	0	0	...
I	...	0	0	0	0.2	0	0		0	0.2	0	0	...
K	...	0	0.2	0.6	0	0	0		0	0	0	0.2	...
L	...	0	0	0	0	0	0		0	0	0	0	...
M	...	0	0	0	0	0	0		0	0	0	0	...
N	...	0	0	0	0	0.25	0		0.6	0	0	0	...
P	...	0	0	0	0	0	0		0	0	0.2	0	...
Q	...	0	0	0	0	0	0		0	0	0	0	...
R	...	0	0	0	0	0	0		0	0	0	0	...
S	...	0	0	0	0	0	0		0	0	0	0	...
T	...	0	0	0	0	0	0		0	0	0	0	...
V	...	0	0	0	0.6	0	0		0	0.4	0	0	...
W	...	1.0	0	0	0	0	0		0	0	0	0	...
Y	...	0	0	0	0	0	0		0	0	0	0	...

HMMs include position-specific gap penalties

Match *or* Delete

Deletions

Insertions

		M/D	M/D	M/D	I	I	M/D	M/D	M/D	M/D	M/D	
HBA_human	. . .	V	G	A	.	.	H	A	G	E	Y	. . .
HBB_human	. . .	V	–	–	.	.	N	V	D	E	V	. . .
MYG_phyca	. . .	V	E	A	.	.	D	V	A	G	H	. . .
LGB2_luplu	. . .	F	N	A	.	.	N	I	P	K	H	. . .
GLB1_glydi	. . .	I	A	G	a	d	N	G	A	G	V	. . .

Probabilities for

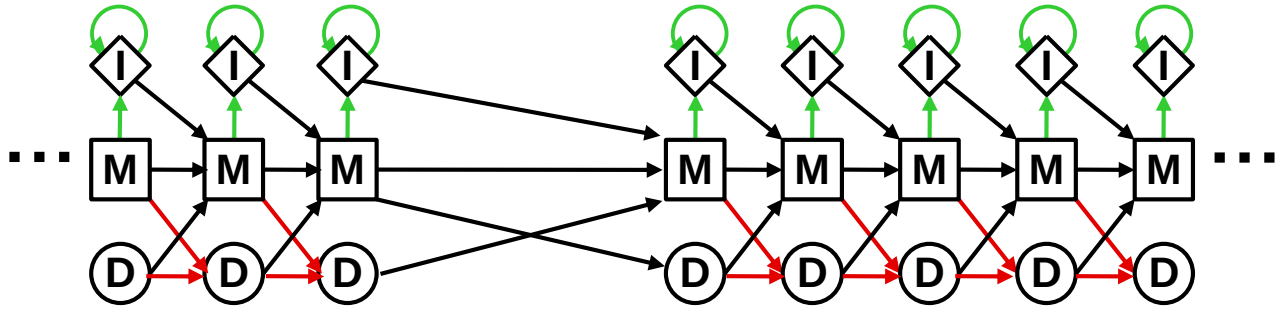
- Insert Open
- Insert Extend
- Delete Open
- Delete Extend

A	. . .	0	0.25	0.75		0	0.2	0.4	0	0	. . .
C	. . .	0	0	0		0	0	0	0	0	. . .
D	. . .	0	0	0		0.2	0	0.2	0	0	. . .
E	. . .	0	0.25	0		0	0	0	0.4	0	. . .
F	. . .	0.2	0	0		0	0	0	0	0	. . .
G	. . .	0	0.25	0.25		0	0.2	0.2	0.4	0	. . .
H	. . .	0	0	0		0.2	0	0	0	0.4	. . .
I	. . .	0.2	0	0		0	0.2	0	0	0	. . .
K	. . .	0	0	0		0	0	0	0.2	0	. . .
L	. . .	0	0	0		0	0	0	0	0	. . .
M	. . .	0	0	0		0	0	0	0	0	. . .
N	. . .	0	0.25	0		0.6	0	0	0	0	. . .
P	. . .	0	0	0		0	0	0.2	0	0	. . .
...											
W	. . .	0	0	0		0	0	0	0	0	. . .
Y	. . .	0	0	0		0	0	0	0	0.2	. . .
M→ I	. . .	0	0	0.25		0	0	0	0	0	. . .
I → I	. . .	0	0	0.5		0	0	0	0	0	. . .
M→ D	. . .	0.2	0	0		0	0	0	0	0	. . .
D→ D	. . .	0	1.0	0		0	0	0	0	0	. . .

Profile HMMs can be represented as states connected by transitions

		M/D	M/D	M/D	I	I	M/D	M/D	M/D	M/D	M/D	
HBA_human	...	V	G	A	.	.	H	A	G	E	Y	...
HBB_human	...	V	-	-	.	.	N	V	D	E	V	...
MYG_phyca	...	V	E	A	.	.	D	V	A	G	H	...
LGB2_luplu	...	F	N	A	.	.	N	I	P	K	H	...
GLB1_glydi	...	I	A	G	a	d	N	G	-	G	V	...

HMM p



Matrix:

$p_i(a)$

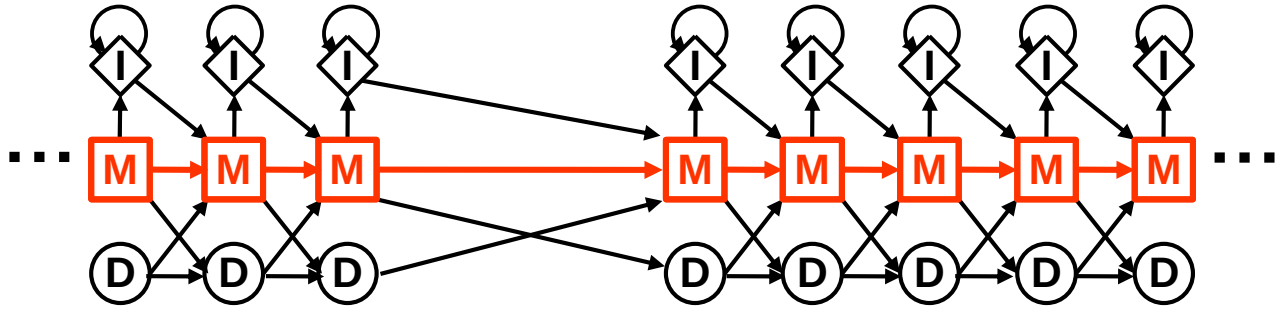
$p_i(X \rightarrow Y)$

A	0	0.25	0.75		0	0.2	0.4	0	0
C	0	0	0		0	0	0	0	0
…									
W	0	0	0		0	0	0	0	0
Y	0	0	0		0	0	0	0	0.2
M→I	0	0	0.25		0	0	0	0	0
I→I	0	0	0.5		0	0	0	0	0
M→D	0.2	0	0		0	0	0	0	0
D→D	0	1.0	0		0	0	0	0	0

Profile HMMs can be represented as states connected by transitions

	M/D	M/D	M/D	I	I	M/D	M/D	M/D	M/D	M/D	
HBA_human	... V	G	A	.	.	H	A	G	E	Y	...
HBB_human	... V	-	-	.	.	N	V	D	E	V	...
MYG_phyca	... V	E	A	.	.	D	V	A	G	H	...
LGB2_luplu	... F	N	A	.	.	N	I	P	K	H	...
GLB1_glydi	... I	A	G	a	d	N	G	-	G	V	...

HMM p



Matrix:

$p_i(a)$

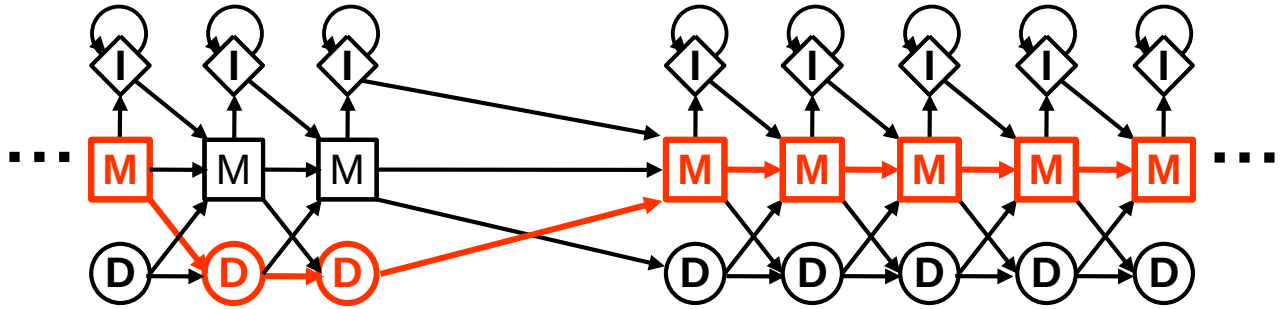
$p_i(X \rightarrow Y)$

A	0	0.25	0.75		0	0.2	0.4	0	0
C	0	0	0		0	0	0	0	0
…									
W	0	0	0		0	0	0	0	0
Y	0	0	0		0	0	0	0	0.2
M→I	0	0	0.25		0	0	0	0	0
I→I	0	0	0.5		0	0	0	0	0
M→D	0.2	0	0		0	0	0	0	0
D→D	0	1.0	0		0	0	0	0	0

Profile HMMs can be represented as states connected by transitions

	M/D	M/D	M/D	I	I	M/D	M/D	M/D	M/D	M/D	
HBA_human	... V	G	A	.	.	H	A	G	E	Y	...
HBB_human	... V	–	–	.	.	N	V	D	E	V	...
MYG_phyca	... V	E	A	.	.	D	V	A	G	H	...
LGB2_luplu	... F	N	A	.	.	N	I	P	K	H	...
GLB1_glydi	... I	A	G	a	d	N	G	–	G	V	...

HMM p



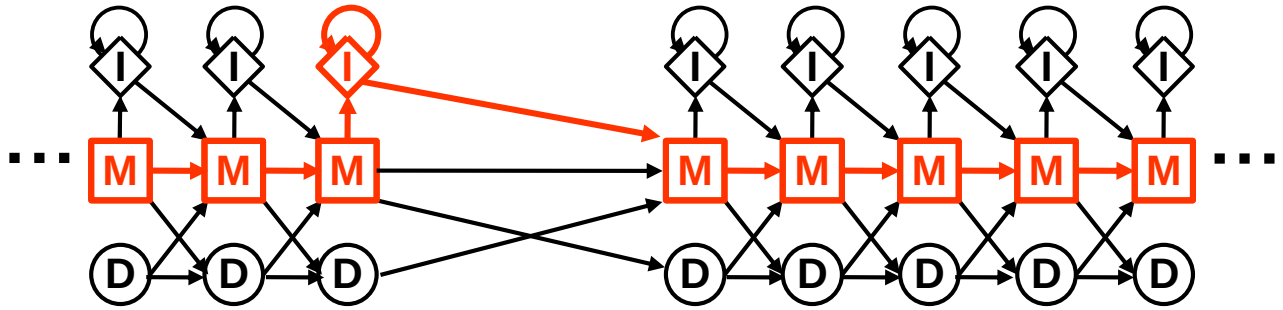
Matrix:

$p_i(a)$	A	0	0.25	0.75		0	0.2	0.4	0	0
	C	0	0	0		0	0	0	0	0
	...									
	W	0	0	0		0	0	0	0	0
	Y	0	0	0		0	0	0	0	0.2
$p_i(X \rightarrow Y)$	M→I	0	0	0.25		0	0	0	0	0
	I→I	0	0	0.5		0	0	0	0	0
	M→D	**0.2**	0	0		0	0	0	0	0
	D→D	0	**1.0**	0		0	0	0	0	0

Profile HMMs can be represented as states connected by transitions

		M/D	M/D	M/D	I	I	M/D	M/D	M/D	M/D	M/D	
HBA_human	...	V	G	A	.	.	H	A	G	E	Y	...
HBB_human	...	V	-	-	.	.	N	V	D	E	V	...
MYG_phyca	...	V	E	A	.	.	D	V	A	G	H	...
LGB2_luplu	...	F	N	A	.	.	N	I	P	K	H	...
GLB1_glydi	...	I	A	G	a	d	N	G	-	G	V	...

HMM p



Matrix:

$p_i(a)$

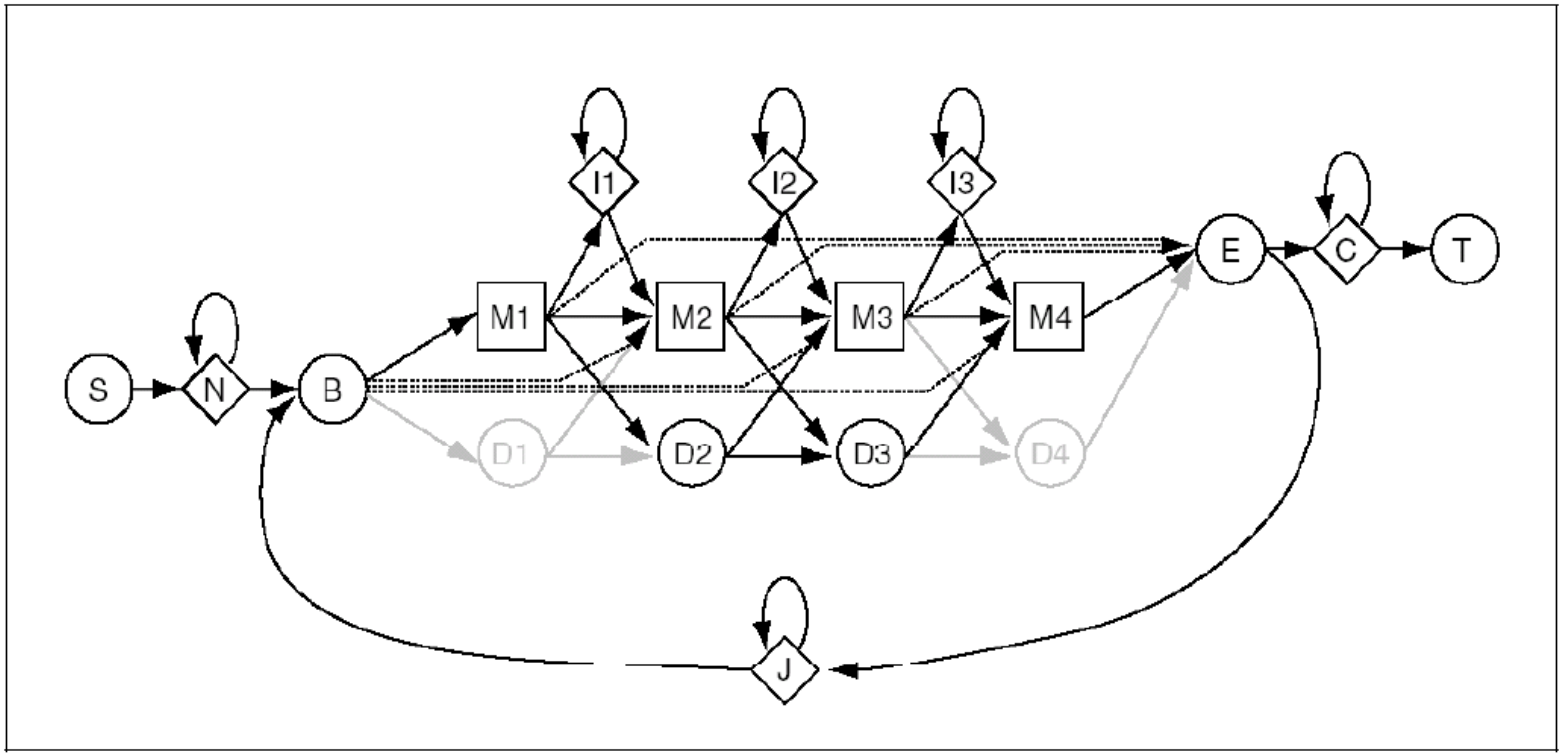
$p_i(X \rightarrow Y)$

A	0	0.25	0.75		0	0.2	0.4	0	0
C	0	0	0		0	0	0	0	0
…									
W	0	0	0		0	0	0	0	0
Y	0	0	0		0	0	0	0	0.2
M→I	0	0	**0.25**		0	0	0	0	0
I→I	0	0	**0.5**		0	0	0	0	0
M→D	0.2	0	0		0	0	0	0	0
D→D	0	1.0	0		0	0	0	0	0

HMMER

- GNU license
- Ανανεώνεται συνεχώς
- Τρέχει σε όλες τις πλατφόρμες
- Ποικιλία εργαλείων
- Ευέλικτη αρχιτεκτονική

HMMER



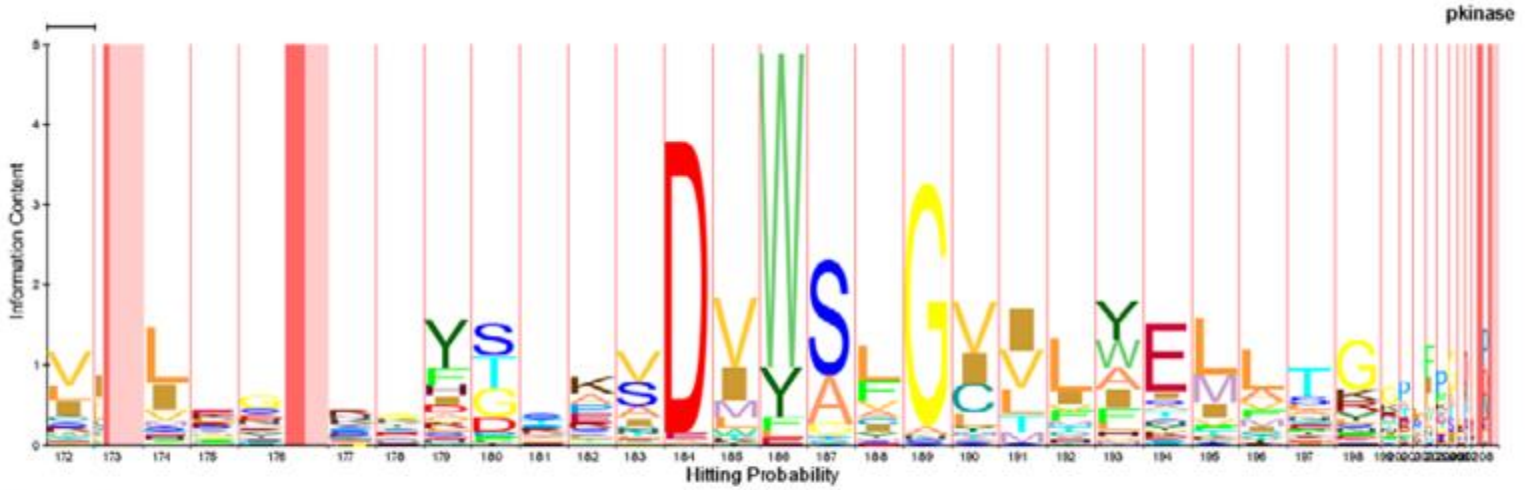


Figure 2
Partial logo (positions 172–209) of the Pfam pkinase model. Positions with narrow match state stacks are likely to be deleted in typical family members. The total width of a red-shaded (dark+light) stack visualizes the expected number of inserted letters. The left dark-shaded part of the stack's width represents the probability that at least one letter is inserted. The difference is illustrated by comparing I_{173} with I_{176}: Both states have approximately the same expected contribution, but the hitting probability of I_{176} is higher. The insertion stack height is zero for all shown examples because the emission probabilities correspond to the background frequencies.

http://biotech.szbk.u-szeged.hu/bioinf/AT-hookHMM-Logo.htm

Ρουτίνες του HMMER

hmmbuild: Πρόγραμμα με χρήση του οποίου, ξεκινώντας από μια αρχική πολλαπλή στοίχιση, κατασκευάζεται ένα μοντέλο HMM το οποίο να την περιγράφει.

hmmalign: Πρόγραμμα με το οποίο μια σειρά ακολουθιών οι οποίες προέρχονται από ένα HMM, στοιχίζονται σε μια πολλαπλή στοίχιση. Η πολλαπλή στοίχιση, επιτυγχάνεται μέσω διαδοχικών στοιχίσεων των ακολουθιών με το μοντέλο.

hmmsearch: Πρόγραμμα το οποίο, πραγματοποιεί αναζητήσεις ενός μοντέλου HMM έναντι μιας βάσης ακολουθιών πρωτεϊνών.

phmmer: Πρόγραμμα το οποίο πραγματοποιεί αναζήτηση μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το BLASTP)

jackhmmer: Πρόγραμμα το οποίο πραγματοποιεί επαναληπτικές αναζητήσεις μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το PSI-BLAST)

Ρουτίνες του HMMER

hmmscan: Πρόγραμμα με το οποίο πραγματοποιούνται αναζητήσεις μιας η περισσότερων ακολουθιών έναντι μιας βάσης δεδομένων από μοντέλα HMM. Πρέπει να τονιστεί εδώ, ότι αν έχουμε μια ακολουθία και ένα HMM, τα δυο παραπάνω προγράμματα επιστρέφουν ακριβώς το ίδιο αποτέλεσμα. Αν διαφέρουν, είτε οι ακολουθίες είτε τα μοντέλα, τότε δίνουν άλλο αποτέλεσμα, λόγω του διαφορετικού τρόπου υπολογισμού της στατιστικής σημαντικότητας.

nhmmer: Πρόγραμμα που πραγματοποιεί αναζήτηση μιας ακολουθίας DNA, μιας στοίχισης ή ενός pHMM, έναντι μιας βάσης ακολουθιών DNA. (ανάλογο με το BLASTN)

nhmmscan: Πρόγραμμα που πραγματοποιεί αναζήτηση μιας ακολουθίας DNA έναντι μιας βάσης δεδομένων από DNA profile HMM.

hmmconvert: Πρόγραμμα που μετατρέπει μοντέλα HMM από και προς τη μορφή του HMMER3.

hmmemit: Πρόγραμμα, με το οποίο 'εκπέμπεται' η καλύτερη (ανάλογα με τον ορισμό) ακολουθία η οποία θα μπορούσε να παραχθεί από το μοντέλο.

hmmpress: Μετατρέπει μια βάση δεδομένων HMM σε δυαδικό κώδικα για το hmmscan.

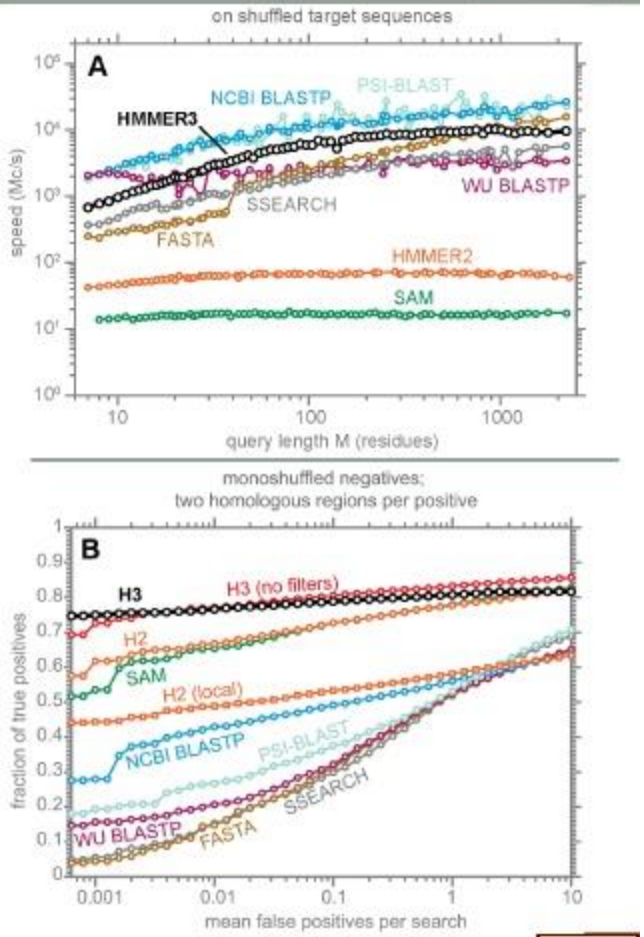
hmmstat: δείχνει συνοπτικά στατιστικά για μια βάση δεδομένων HMM.

Πλεονεκτήματα του HMMER

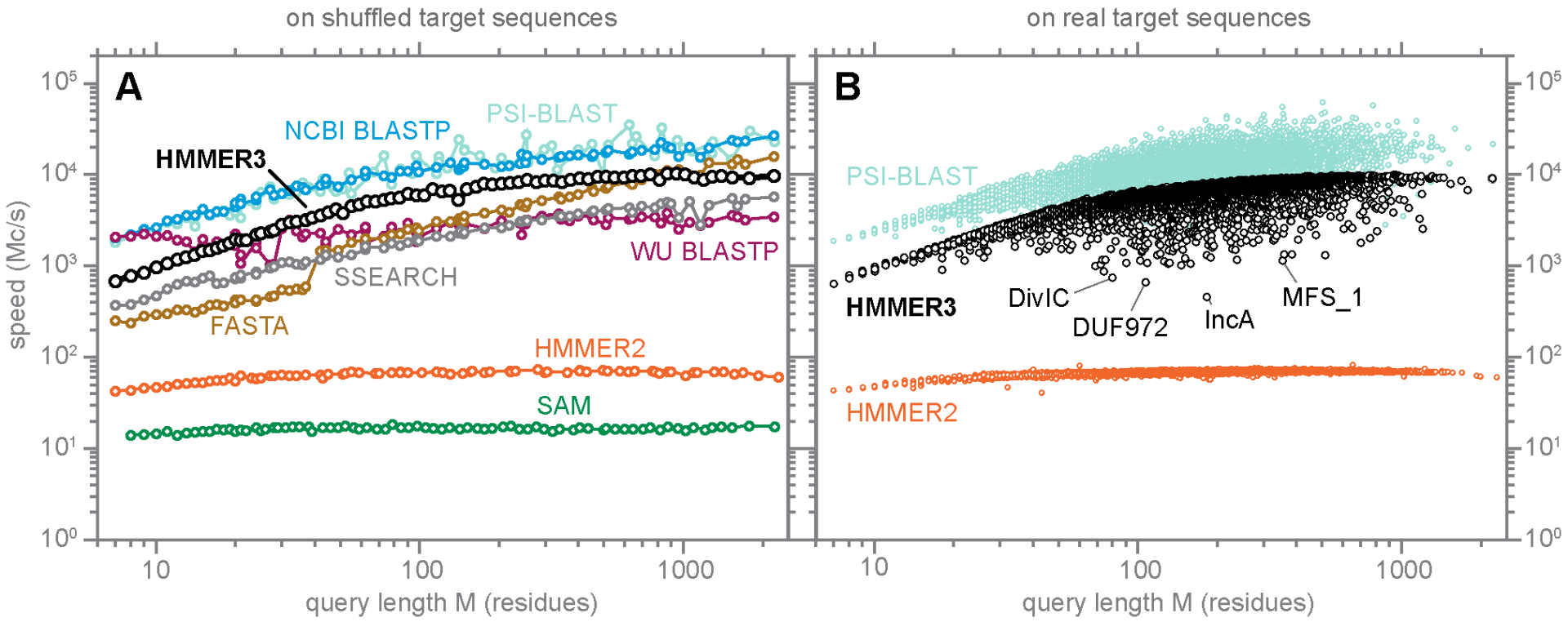
- Ιδιαίτερα εύκολο, τόσο στην κατασκευή μοντέλου όσο και σε απλές αναζητήσεις σε βάσεις δεδομένων
- Άριστη μαθηματική θεμελίωση της στοίχισης, δεν υπάρχει καμία αυθαιρεσία (πχ ποινές για τα κενά κλπ)
- Με την έκδοση 3.0 παρέχει και δυνατότητα για απευθείας αναζήτηση μιας αλληλουχίας σε μια βάση δεδομένων αλληλουχιών με όμοιο τρόπο με το BLAST, PSI-BLAST (κατασκευάζει «στον αέρα» ένα μοντέλο HMM από την αλληλουχία επερώτησης και το ανανεώνει με τα αποτελέσματα της αναζήτησης)
- Ταχύτητα που συγκρίνεται πλέον με αυτή του BLAST
- Εύκολος υπολογισμός της στατιστικής σημαντικότητας (E-value)

HMMER vs BLAST

	HMMER	BLAST
Program	*PHMMER*	*BLASTP*
Query	Single sequence	
Target Database	Sequence database	
Program	*HMMSCAN*	*RPSBLAST*
Query	Single sequence	
Target Database	Profile HMM database, e.g. Pfam	PSSM database, e.g. CDD
Program	*HMMSEARCH*	*PSI-BLAST*
Query	Profile HMM	PSSM
Target Database	Sequence database	
Program	*JACKHMMER*	*PSI-BLAST*
Query	Single sequence	
Target Database	Sequence database	



Modified from: S. R. Eddy
PLoS Comp. Biol., 7:e1002195, 2011.



Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7(10): e1002195. https://doi.org/10.1371/journal.pcbi.1002195

SAM

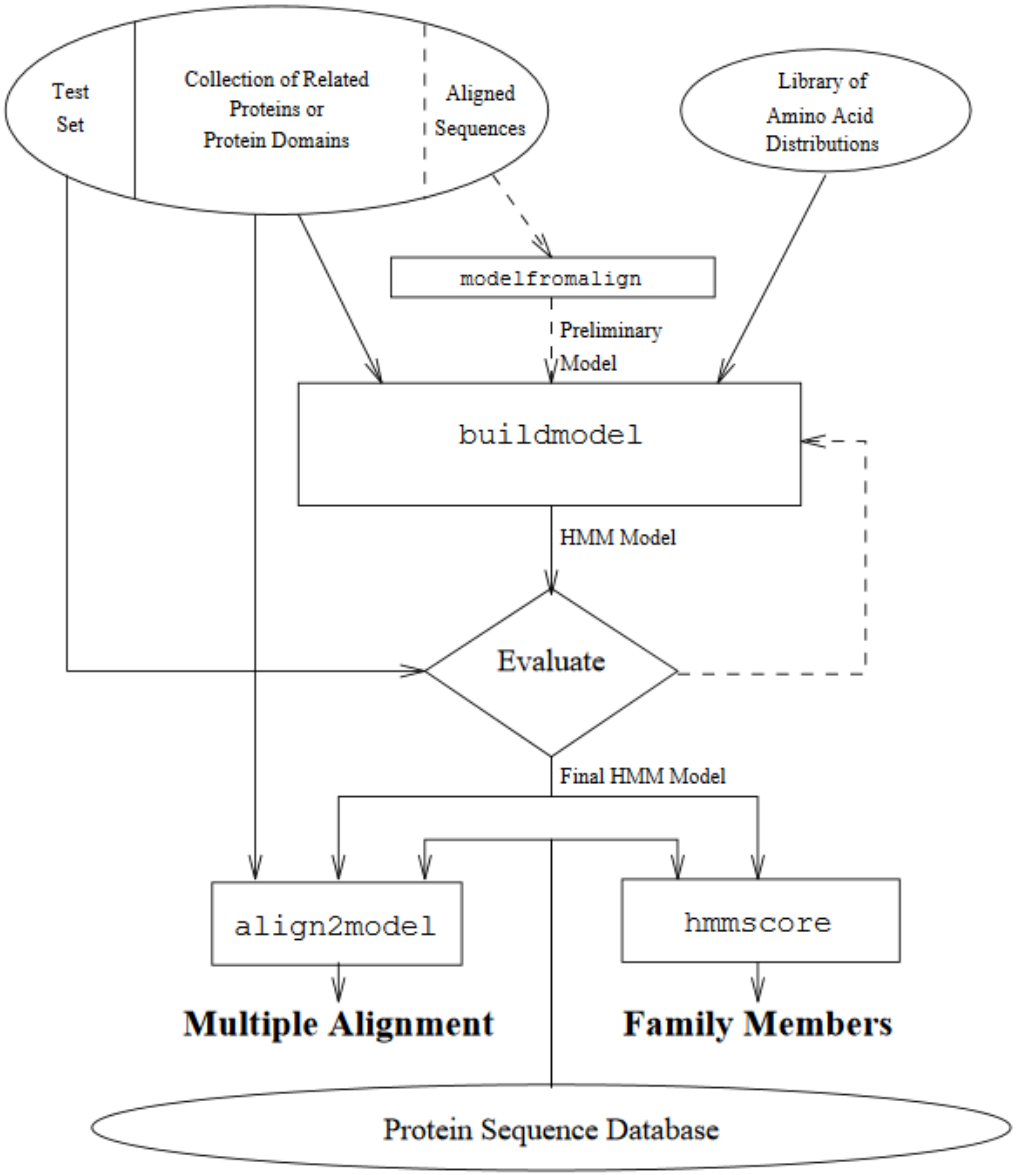
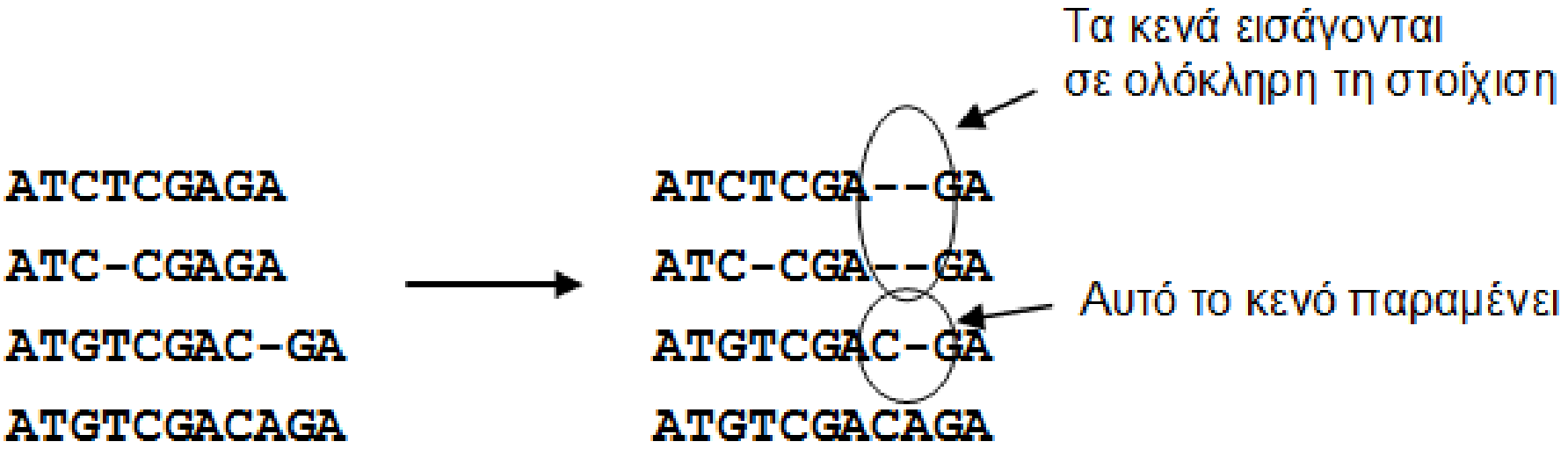


Figure 2: Overview of SAM.

Μέθοδοι profile-profile alignment

Στις μεθόδους αυτές περιλαμβάνονται οι μεθοδολογίες που στοιχίζουν μεταξύ τους προφίλ αλληλουχιών ή PSSM

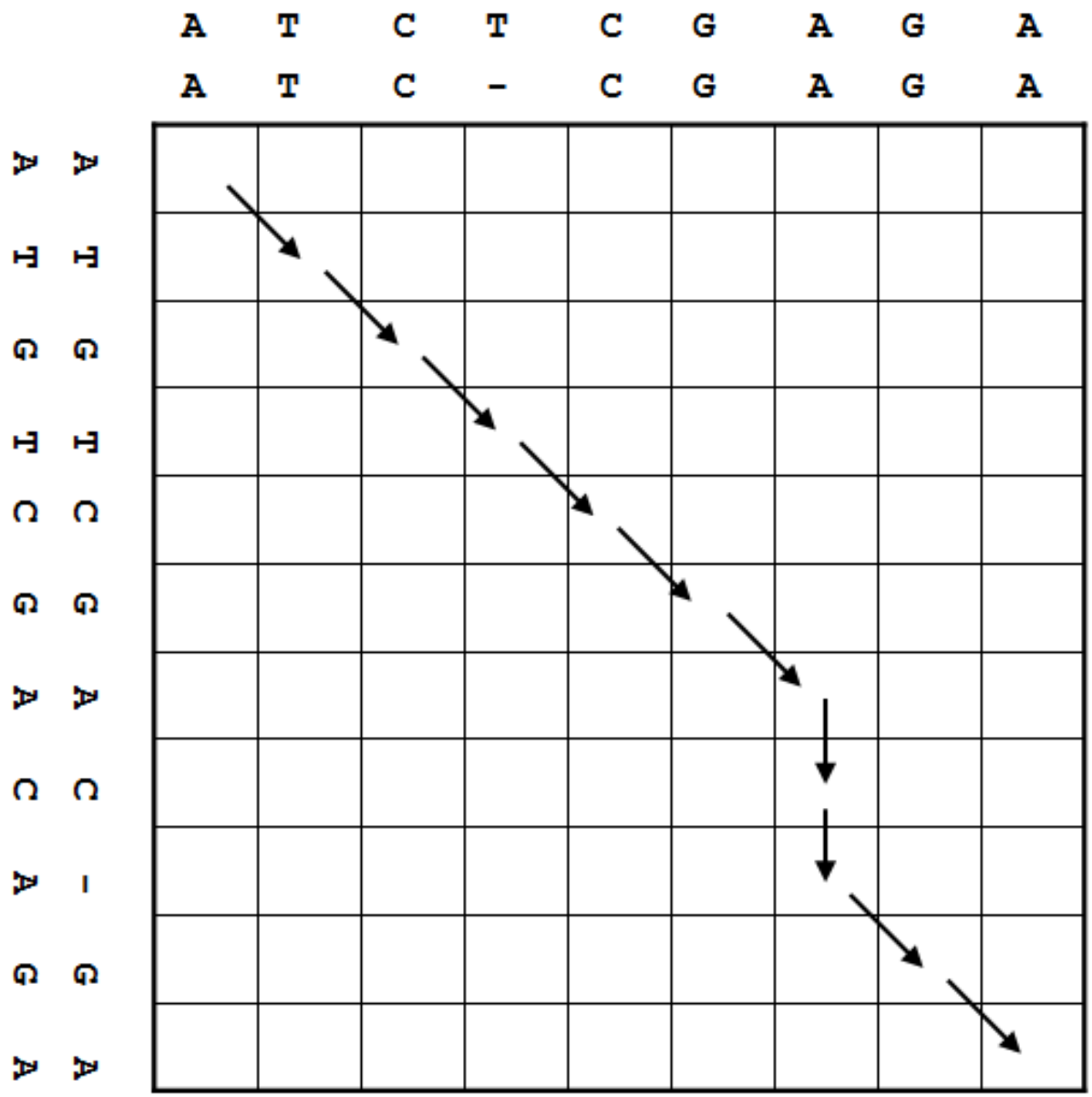
- FFAS (Fold and Function Assignment System)
- PROF_SIM
- COMPASS (comparison of multiple protein alignments with assessment of statistical significance)
- COMA (Comparison Of Multiple Alignments)



Το λεγόμενο *profile alignment*, μετράει τη σχετική συνεισφορά όλων των ακολουθιών της κάθε στοίχισης και τελικά πραγματοποιεί την στοίχιση λαμβάνοντας υπόψη όλες τις ακολουθίες. Τα μαθηματικά της μεθόδου είναι πολύπλοκα, αλλά μπορούν να απλοποιηθούν αν θεωρήσουμε, όπως και παραπάνω, το κενό σαν ένα πέμπτο σύμβολο (-), οπότε θα έχουμε και γραμμική ποινή για τα κενά

$$\sum_i SP(m_i) = \sum_i \sum_{j<j'} s\left(m_i^j, m_i^{j'}\right)$$

$$= \sum_i \sum_{j<j'\le n} s\left(m_i^j, m_i^{j'}\right) + \sum_i \sum_{n<j<j'\le N} s\left(m_i^j, m_i^{j'}\right) + \sum_i \sum_{j\le n, n<j'\le N} s\left(m_i^j, m_i^{j'}\right)$$


A T C T C G A G A
A T C - C G A G A
A T G T C G A C A G A
A T G T C G A C - G A

Γενικά

- Σε γενικές γραμμές οι μέθοδοι αυτές χρησιμοποιούν παραλλαγές του αλγορίθμου δυναμικού προγραμματισμού για να στοιχίσουν με βέλτιστο τρόπο δύο προφίλ
- Οι διαφοροποιήσεις μεταξύ τους εντοπίζονται κυρίως
  - Στο είδος του προφίλ (PSI-BLAST, PSSM, Weight Matrix)
  - Στον τρόπο που υπολογίζουν το σκορ (Dot product, Sum of Pairs etc)
  - Στον τρόπο υπολογισμού της ποινής για τα κενά
  - Στον τρόπο εύρεσης της στατιστικής σημαντικότητας
  - Στον τρόπο ενσωμάτωσης δομικής πληροφορίας

Scoring

- Sum of pairs

$$S_{1,2} = \sum_{i=1}^{20} \sum_{j=1}^{20} Q_i^1 Q_j^2 s_{ij}$$

- Log-average

$$S_{1,2} = \ln \sum_{a=1}^{20} \sum_{b=1}^{20} Q_a^1 Q_b^2 q_{ab}$$

- Dot product

$$S_{1,2} = \sum_{i=1}^{20} Q_i^1 Q_i^2$$

- Pearson Correlation

$$S = \frac{\sum_{i=1}^{20} (W_i^1 - \langle W_i^1 \rangle)(W_i^2 - \langle W_i^2 \rangle)}{\sqrt{\sum_{i=1}^{20} (W_i^1 - \langle W_i^1 \rangle)^2 \sum_{i=1}^{20} (W_i^2 - \langle W_i^2 \rangle)^2}}$$

- Log-odds

$$S_{1,2} = \sum_{i=1}^{20} F_i^1 \ln \frac{Q_i^2}{p_i} \qquad S_{1,2} = \sum_{i=1}^{20} F_i^1 \ln \frac{Q_i^2}{p_i} + \sum_{i=1}^{20} F_i^2 \ln \frac{Q_i^1}{p_i}$$

- Information Theoretic

$$S = \frac{1}{2}(1 - D)(1 + S) \qquad S_{1,2} = c_1 S_1 + c_2 S_2$$

$$S = \frac{1}{2}\left[\sum_{i=1}^{20} Q_i^0 \log_2 \frac{Q_i^0}{R_i^0} + \sum_{i=1}^{20} p_i \log_2 \frac{p_i}{R_i^0}\right]$$

$$D = \frac{1}{2}\left[\sum_{i=1}^{20} Q_i^1 \log_2 \frac{Q_i^1}{Q_i^0} + \sum_{i=1}^{20} Q_i^2 \log_2 \frac{Q_i^2}{Q_i^0}\right]$$

$$S_1 = \sum_{i=1}^{20} N_i^1 \ln \frac{Q_i^2}{p_i}; \qquad S_2 = \sum_{i=1}^{20} N_i^2 \ln \frac{Q_i^1}{p_i}$$

Συγκρίσεις των διαφόρων score functions

- Έχουν πραγματοποιηθεί αρκετές συγκριτικές μελέτες για την αξιολόγηση της αποδοτικότητας των διαφορετικών συναρτήσεων.
- Γενικά η σύγκριση είναι δύσκολη γιατί πρέπει να αξιολογηθεί μόνο αυτή η παράμετρος, ανεξάρτητα από τα υπόλοιπα χαρακτηριστικά των μεθόδων.
- Τα γενικά συμπεράσματα συγκλίνουν στο ότι οι διαφορές είναι μικρές, αλλά οι συναρτήσεις που βασίζονται στα log-odds και τη θεωρία της πληροφορίας έχουν κάπως καλύτερη απόδοση σε σχέση με τα dot products, average κλπ

Ohlson T, Wallner B, Elofsson A. Profile–profile methods provide improved fold-recognition: A study of different profile–profile alignment methods. Proteins: Structure, Function, and Bioinformatics. 2004 Oct 1;57(1):188-97.

Wang G, Dunbrack Jr RL. Scoring profile-to-profile sequence alignments. Protein Science. 2004 Jun;13(6):1612-26.

Edgar RC, Sjölander K. A comparison of scoring functions for protein sequence profile alignment. Bioinformatics. 2004 Feb 12;20(8):1301-8.

FFAS

Κάθε μέθοδος στοίχισης προφίλ κάθε προφίλ περιλαμβάνει τέσσερα βήματα:

- (i) την προετοιμασία της πολλαπλής στοίχισης ακολουθιών,
- (ii) την κατασκευή του προφίλ,
- (iii) την στοίχιση του προφίλ με τα προφίλ αλληλουχιών από τη βάση δεδομένων, όπως η PDB, και
- (iv) τη στατιστική σημασία της βαθμολογίας ευθυγράμμισης.

Στη μέθοδο FFAS, η πολλαπλή στοίχιση αλληλουχιών κατασκευάζεται χρησιμοποιώντας το PSI-BLAST. Πραγματοποιούνται πέντε επαναλήψεις με αναζήτηση του PSI-BLAST έναντι της βάσης δεδομένων NR85S των πρωτεϊνικών αλληλουχιών. Στο δεύτερο βήμα, όλες οι ακολουθίες που εντοπίστηκαν από το PSI-BLAST με τιμή E <0,005 χρησιμοποιούνται για την κατασκευή του προφίλ. Τα βάρη αποδίδονται σε αλληλουχίες βάσει της ομοιότητάς τους με άλλες αλληλουχίες στην πολλαπλή στοίχιση

Η τιμή του σκορ της σύγκρισης μεταξύ των θέσεων *η* και *m* στα δύο προφίλ υπολογίζεται ως το εσωτερικό γινόμενο της στήλης *n* από το πρώτο προφίλ και της στήλη *m* από το δεύτερο προφίλ. Μετά την εκχώρηση τιμών σε όλες τις θέσεις, ο πίνακας κανονικοποιείται. Η βέλτιστη στοίχιση υπολογίζεται από έναν αλγόριθμο δυναμικού προγραμματισμού. Στο τελευταίο βήμα, το καθαρό σκορ της στοίχισης μεταφράζεται στο τελικό σκορ του FFAS συγκρίνοντάς το με την κατανομή των σκορ που λαμβάνονται για ζεύγη μη σχετιζόμενων πρωτεϊνών. Η τρέχουσα έκδοση βασίζεται στην ίδια προσέγγιση με μικρές τροποποιήσεις στο σύστημα σύγκρισης προφίλ και βαθμολόγησης.

Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W284-8.

FFAS

Σε μια μεταγενέστερη δημοσίευση, η διαδικτυακή εφαρμογή εμπλουτίστηκε σε διάφορα σημεία. Η βάση δεδομένων που χρησιμοποιήθηκε για τον υπολογισμό των προφίλ εμπλουτίστηκε με την προσθήκη δημόσια διαθέσιμων μεταγονιδιωματικών αλληλουχιών. Το προφίλ της πρωτεΐνης που υποβάλλει ένας χρήστης μπορεί πλέον να συγκριθεί με διάφορες βάσεις δεδομένων, συμπεριλαμβανομένων αρκετών πλήρως προσδιορισμένων πρωτεομάτων, ανθρώπινων πρωτεϊνών που εμπλέκονται σε γενετικές ασθένειες και δεδομένων μικροβιακών παραγόντων μολυσματικότητας. Η νέα διεπαφή χρησιμοποιεί ένα σύστημα καρτελών, επιτρέποντας στον χρήστη να πλοηγεί σε πολλαπλές σελίδες και περιλαμβάνει επίσης νέες λειτουργίες, όπως ένα dotplot graph viewer, εργαλεία μοντελοποίησης, βελτιωμένο πρόγραμμα οπτικοποίησης στοίχισης δομών και συνδέσμους σε άλλες βάσεις δεδομένων. Ο διακομιστής FFAS βελτιστοποιήθηκε επίσης για ταχύτητα: οι χρόνοι λειτουργίας μειώθηκαν κατά μία τάξη μεγέθους. Ο διακομιστής FFAS, http://ffas.godziklab.org, δεν έχει απαίτηση σύνδεσης χρήστη, αν και υπάρχει η δυνατότητα εγγραφής και αποθήκευσης των αποτελεσμάτων σε μεμονωμένους καταλόγους που προστατεύονται με κωδικό πρόσβασης. Ο πηγαίος κώδικας και τα εκτελέσιμα Linux για το πρόγραμμα FFAS είναι διαθέσιμα για λήψη από το διακομιστή FFAS.

Jaroszewski L, Li Z, Cai XH, Weber C, Godzik A. FFAS server: novel features and applications. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W38-44. doi: 10.1093/nar/gkr441.

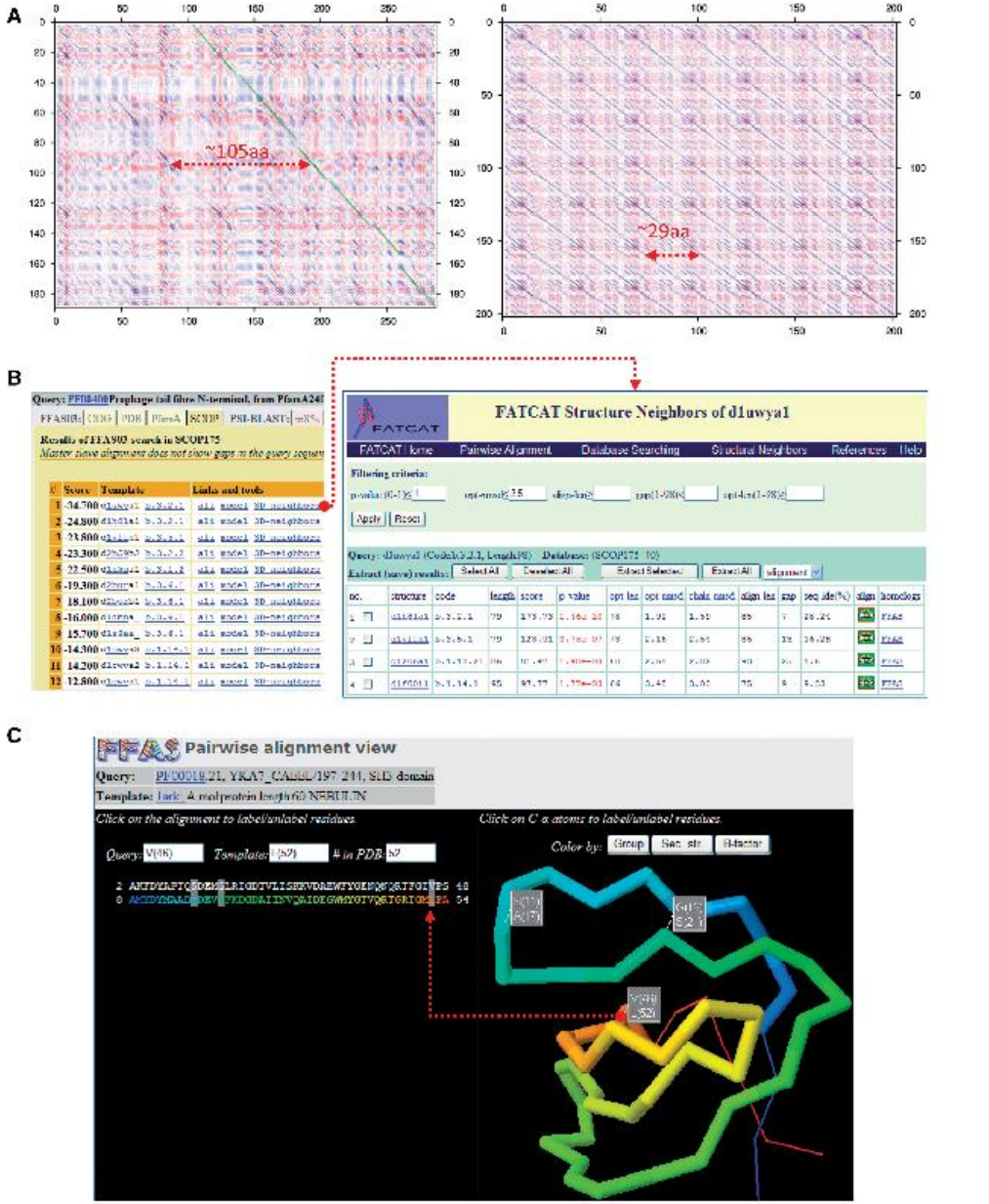


Figure 1. Examples of novel features of the FFAS server. **(A)** Dotplot graphs generated with the new FFAS tool. Left panel: the dotplot graph of a leucine-rich repeat region of the human NACHT protein compared to itself. Right panel: the dotplot graph visualizing similarity between C-terminal parts of SusE and SusF proteins from *Bacteroides thethaiotaomicron*. Arrows indicate the estimated lengths of repeats in NACHT LRRs and the lengths of repeated (homologous) domains in the alignment of SusE with SusF. **(B)** FFAS results are now linked to a database of structural similarities calculated with FATCAT. These links can be used to evaluate structural consistency of FFAS results. In this example, the fact that two different folds are aligned with the same query (Prophage tail fibre N-terminal domain) is explained by a list of structural neighbors that shows that a Prealbumin-like fold (b.3 code in SCOP) and an Immunoglobulin-like beta-sandwich (b.1 code in SCOP) are structurally similar despite being classified as separate folds. **(C)** 3D alignment viewer allows quick inspection of the alignment as 'projected' on a template structure (labeling of residues in a Jmol viewer is synchronized with alignment labeling).

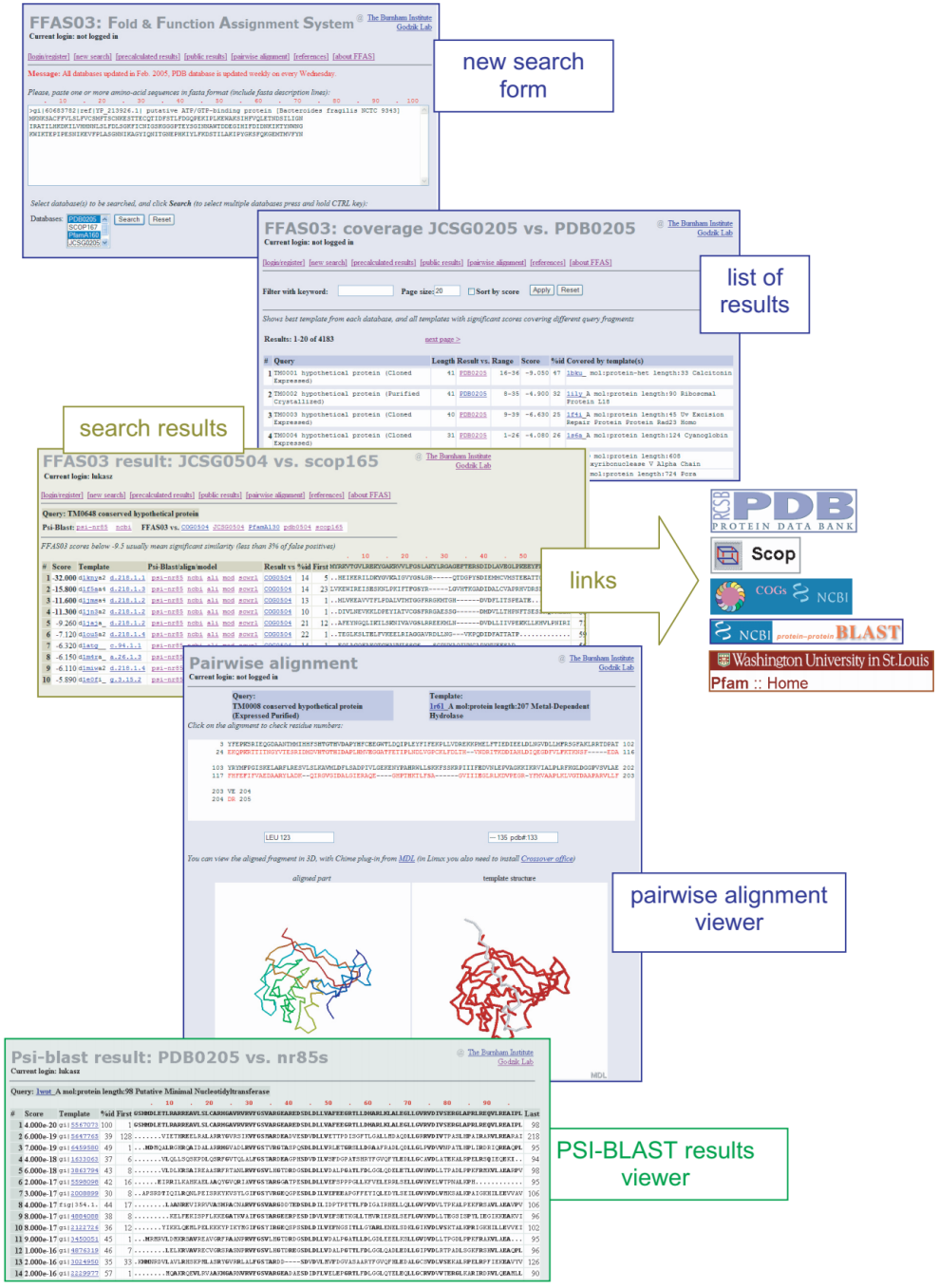


Figure 1. The overview of FFAS pages.

PROF_SIM

Το PROF_SIM είναι μια από τις πιο παλιές μεθόδους. Η μέθοδος συγκρίνει δύο προφίλ εισόδου (όπως αυτά που παράγονται από το PSI-BLAST) και αποδίδει ένα σκορ ομοιότητας για να αξιολογήσει την στατιστική τους ομοιότητα. Η μέθοδος σύγκρισης προφίλ-προφίλ, η οποία επιτρέπει την εισαγωγή κενών, μπορεί να χρησιμοποιηθεί για την ανίχνευση μακρινών ομοιοτήτων μεταξύ των οικογενειών πρωτεϊνών. Έχει επίσης βελτιστοποιηθεί για να παράγει στοιχίσεις που είναι σε πολύ καλή συμφωνία με τις δομικές στοιχίσεις. Δοκιμές έδειξαν ότι οι στοιχίσεις προφίλ-προφίλ σχετίζονται πράγματι σε μεγάλο βαθμό με τις ομοιότητες μεταξύ των στοιχείων δευτεροταγούς δομής και της τριτοταγούς δομής. Αξιολογήσεις έδειξαν ότι η μέθοδος είναι σημαντικά πιο ευαίσθητη στην ανίχνευση απομακρυσμένων ομολογιών από τα δημοφιλή προγράμματα αναζήτησης PSI-BLAST και IMPALA. Η σχετική βελτίωση είναι της ίδιας τάξης μεγέθους με τη βελτίωση που επιτυγχάνει το PSI-BLAST έναντι του απλού BLAST. Το PROF_SIM συχνά εντοπίζει ομοιότητες που εμπίπτουν στη ζώνη του λυκόφωτος (<30%)

Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol. 2002 Feb 1;315(5):1257-75.

PROF_SIM

Η σύγκριση προφίλ-προφίλ εκτελείται χρησιμοποιώντας αλγόριθμο δυναμικού προγραμματισμού και η στοίχιση λαμβάνει ένα σκορ που αντιστοιχεί σε ταυτίσεις, εισαγωγές και απαλοιφές, με τον ίδιο τρόπο που υπολογίζεται στην περίπτωση απλής στοίχισης αλληλουχιών. Οι διαφορές εντοπίζονται στον τρόπο με τον οποίο γίνεται η βαθμονόμηση και ο υπολογισμός του σκορ. Σε αντίθεση με την απλή σύγκριση αλληλουχίας-αλληλουχίας, όπου ένας πίνακας ομοιότητας όπως ο BLOSUM62 δίνει τη το σκορ για διαφορετικά ζεύγη στοιχισμένων αμινοξέων, η σύγκριση προφίλ-προφίλ είναι πιο περίπλοκη. Ο πυρήνας της διαδικασίας είναι ο ορισμός των σκορ ομοιότητας προφίλ (profile similarity scores) και οι παράμετροι που χρησιμοποιούνται για τον ποσοτικό προσδιορισμό αυτού του μέτρου ομοιότητας.

$$S = \frac{1}{2}(1 - D)(1 + S)$$

$$D = \frac{1}{2}\left[\sum_{i=1}^{20} Q_i^1 \log_2 \frac{Q_i^1}{Q_i^0} + \sum_{i=1}^{20} Q_i^2 \log_2 \frac{Q_i^2}{Q_i^0}\right] \quad S = \frac{1}{2}\left[\sum_{i=1}^{20} Q_i^0 \log_2 \frac{Q_i^0}{R_i^0} + \sum_{i=1}^{20} p_i \log_2 \frac{p_i}{R_i^0}\right]$$

Table 2. Performance evaluation results

Relationship-type	Number of relationships with *E*-value ⩽ 0.1 detected by:			
	Gapped-BLAST	IMPALA	PSI-BLAST	*prof_sim*
Same superfamily (true relationship)	116	115	146	166
Same fold (''possible'' relationship)	0	0	3	1
Same class (''possible'' relationship)	18	14	17	14
Different class (''suspicious'' relationship)	31	20	31	21
Alpha ↔ Beta (''error'' relationship)	1	1	1	0
Total (with *E*-value ⩽ 0.1)	166	150	198	202

For each method, the number of true, possible, suspicious and error relationships that are detected with *E*-value ⩽ 0.1 is reported

Table 3. Performance evaluation results

Relationship-type	Number of relationships detected by:				
	Gapped-BLAST	IMPALA	PSI-BLAST	*prof_sim*	*Structal*
Same superfamily (true relationship)	116	120	146	173	355
Same fold (''possible'' relationship)	0	0	2	1	45
Same class (''possible'' relationship)	18	21	17	18	5
Different class (''suspicious'' relationship)	31	28	30	30	0
Alpha ↔ Beta (''error'' relationship)	1	1	1	1	0
Total	166	170	196	223	405
E-value	0.1	0.14	0.1	0.14	4.93e-07
ROC area	5155	5322 (3.3 %)	6335 (23 %)	7266 (41 %)	13667 (165 %)

For each method, we report the number of true, possible, suspicious and error relationships that are detected until 50 false connections occur (a false connection is everything but a true relationship). Also given are the *E*-value at which the 50th error occurs, and the area under the ROC plot, with the relative improvement with respect to BLAST in parentheses. The last column lists the number of relationships that are detected with *Structal*.

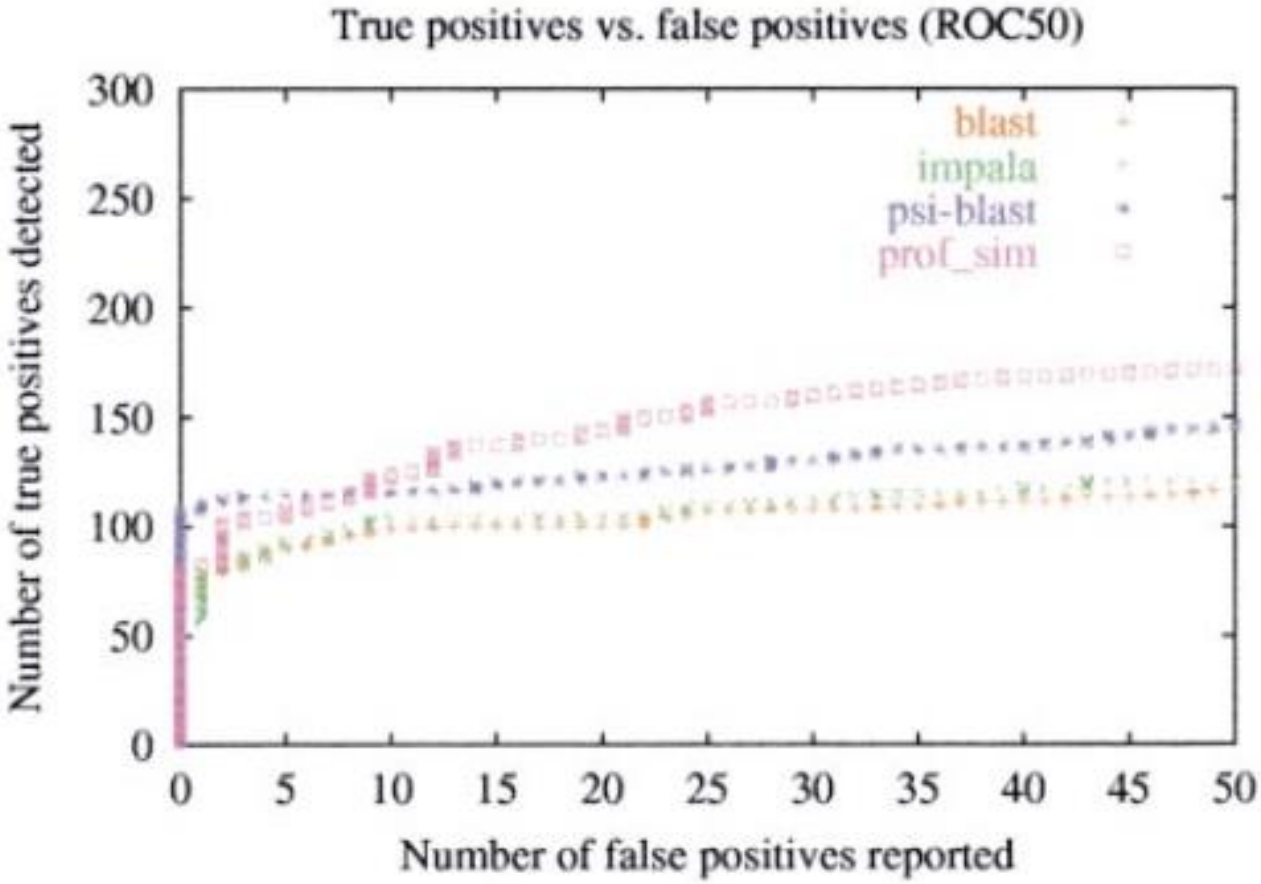


Figure 6. ROC50 curves. A true positive is defined as a connection between families within the same superfamily. Note that the relative improvement of *prof_sim* with respect to PSI-BLAST is comparable to the relative improvement of PSI-BLAST with respect to BLAST (see also Table 3).

COMPASS

Το COMPASS (comparison of multiple protein alignments with assessment of statistical significance) είναι μια άλλη παλιά αλλά ιδιαίτερα αποδοτική μέθοδος. Η μέθοδος παράγει αριθμητικά προφίλ από τις στοιχίσεις, κατασκευάζει τις βέλτιστες τοπικές στοιχίσεις προφίλ-προφίλ και εκτιμά με αναλυτικό τρόπο τις τιμές E (E-values) για τις ανιχνεύσιμες ομοιότητες. Το σύστημα βαθμονόμησης και ο υπολογισμός της τιμής E βασίζονται σε μια γενίκευση της προσέγγισης PSI-BLAST για τη σύγκριση προφίλ-ακολουθίας, η οποία προσαρμόζεται για την περίπτωση στοίχισης προφίλ-προφίλ. Όταν συγκρίθηκε με τις υπάρχουσες μεθόδους σύγκρισης προφίλ-ακολουθίας (PSI-BLAST) και προφίλ-προφίλ (prof_sim), το COMPASS έδειξε αυξημένες ικανότητες για ευαίσθητη και επιλεκτική ανίχνευση απομακρυσμένων ομολόγων, καθώς και βελτιωμένη ποιότητα τοπικών στοιχίσεων. Η μέθοδος επιτρέπει την πρόβλεψη των σχέσεων μεταξύ των οικογενειών πρωτεϊνών στη βάση δεδομένων PFAM πέρα από το εύρος των συμβατικών μεθόδων.

Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol. 2003 Feb 7;326(1):317-36.

Given the effective counts (N) and target frequencies (Q) for every amino acid in the columns from the two compared profiles, two terms are calculated:

$$S_1 = \sum_{i=1}^{20} N_i^1 \ln \frac{Q_i^2}{p_i}; \qquad S_2 = \sum_{i=1}^{20} N_i^2 \ln \frac{Q_i^1}{p_i}$$

where p_i are the background frequencies of the amino acids in the database.

S_1 is a log-odds ratio corresponding to the probability of column 1 to occur given the target frequencies of column 2. S_2 is a log-odds ratio corresponding to the probability of column 2 to occur given the target frequencies of column 1. The two terms S_1 and S_2 are mixed with weights c_1 and c_2 to yield a final substitution score for the given columns 1 and 2:

$$S_{1,2} = c_1 S_1 + c_2 S_2$$

We tested two weighting variants for c_1 and c_2:

$$c_1 = \frac{\sum_{i=1}^{20} N_i^2 - 1}{\sum_{i=1}^{20} N_i^1 + \sum_{i=1}^{20} N_i^2 - 2}$$

$$c_2 = \frac{\sum_{i=1}^{20} N_i^1 - 1}{\sum_{i=1}^{20} N_i^1 + \sum_{i=1}^{20} N_i^2 - 2}$$

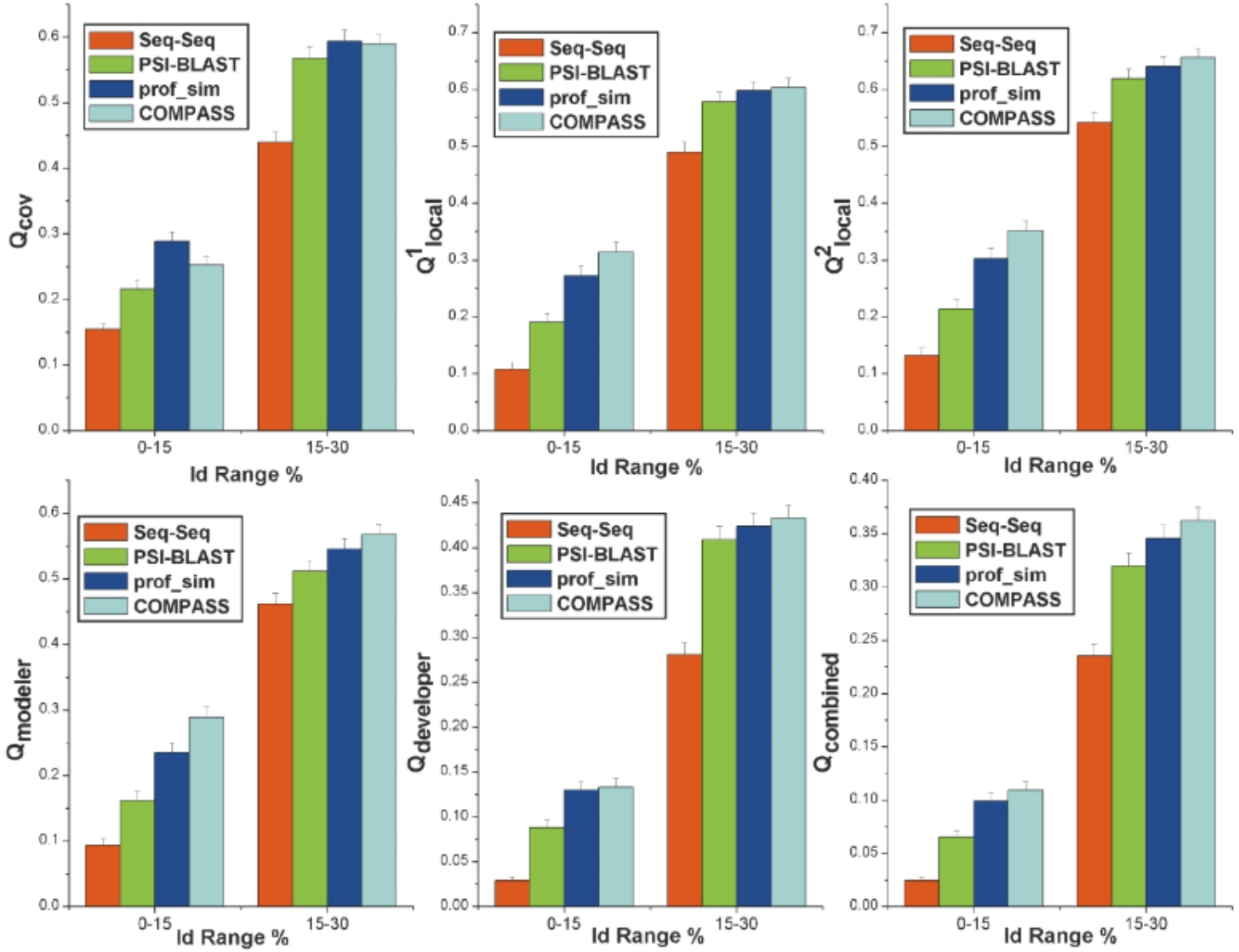


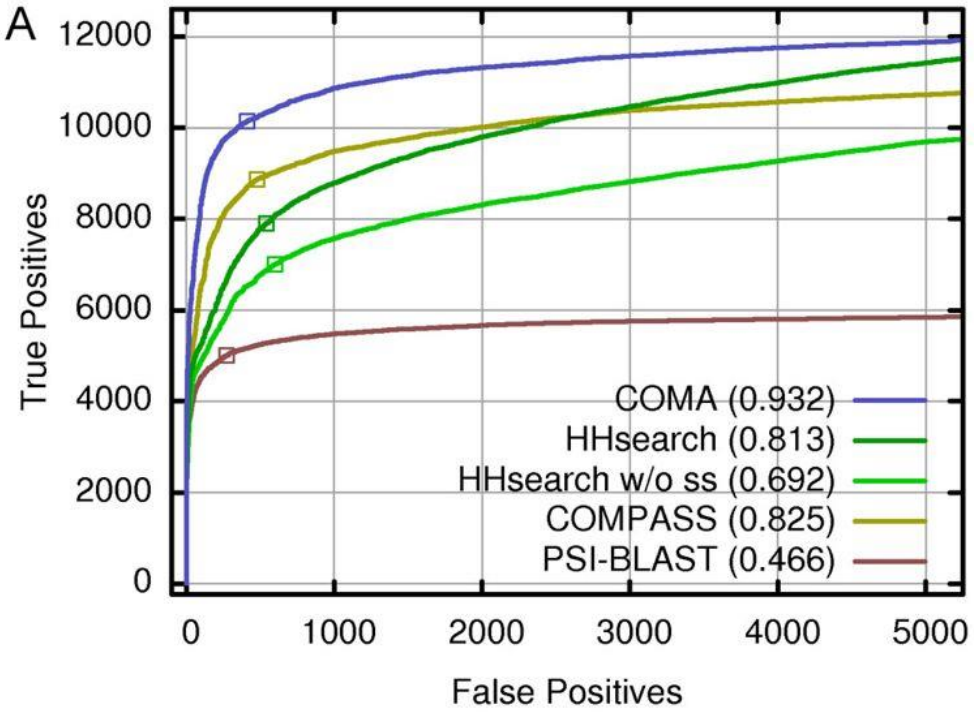
Fig. 3. Evaluation of alignment quality. In two ranges of sequence identity, the quality of the produced local alignments was assessed by six parameters for four different alignment methods (pairwise sequence alignment using BLOSUM62 matrix and Smith-Waterman algorithm, profile–sequence alignment using PSI-BLAST, profile–profile alignments using prof_sim and COMPASS). As a reference, DALI structural alignments from the FSSP database were used. Q_{cov} corresponds to the portion of the length of the structural alignment that was covered by the sequence alignment, regardless of the actual accuracy. Q^1_{local} and Q^2_{local} correspond to the accuracy of the local prediction for only the regions that are included in the evaluated alignment. $Q_{modeler}$, $Q_{developer}$ and $Q_{combined}$ are previously suggested measures of integral accuracy of the alignment from the modeler's, developer's and combined points of view (see the text for details). Means + standard errors are shown.

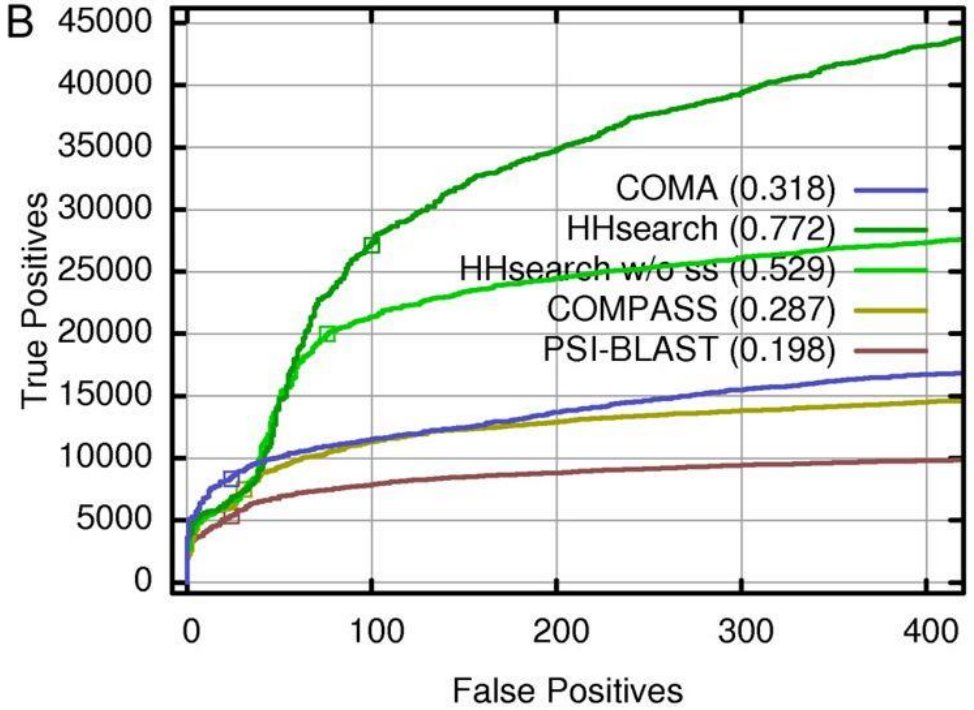
COMA

Η μέθοδος έχει μια σειρά από νέα χαρακτηριστικά, συμπεριλαμβανομένων των ποινών για τα κενά που εξαρτώνται από τη θέση και ενός συστήματος ολικού σκορ. Οι ποινές για τα κενά που εξαρτώνται από τη θέση παρέχουν έναν πιο βιολογικώς αποδεκτό τρόπο για τη στοίχιση οικογενειών πρωτεϊνών ως προφίλ αλληλουχιών. Το σύστημα ολικού σκορ επιτρέπει μια αναλυτική λύση των στατιστικών παραμέτρων που απαιτούνται για την εκτίμηση της στατιστικής σημαντικότητας των ομοιοτήτων προφίλ-προφίλ. Η νέα μέθοδος, μαζί με άλλες σύγχρονες μεθόδους (HHsearch, COMPASS και PSI-BLAST), μελετήθηκε σε μια συγκριτική αξιολόγηση ενός δύσκολου συνόλου δεδομένων της SCOP με πολύ το 20% ταυτότητα αλληλουχίας. Τα αποτελέσματα της αξιολόγησης έδειξαν ότι σε επίπεδο τομέων πρωτεϊνών η μέθοδος συγκρίνεται ευνοϊκά με όλες τις άλλες δοκιμασμένες μεθόδους.

- http://www.bti.vu.lt/en/departments/department-of-bioinformatics/software/coma

Margelevičius, M. and Č. Venclovas (2010). "Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison." BMC Bioinformatics 11(1): 89.


A
True Positives
False Positives
0
2000
4000
6000
8000
10000
12000
0
1000
2000
3000
4000
5000
COMA (0.932)
HHsearch (0.813)
HHsearch w/o ss (0.692)
COMPASS (0.825)
PSI-BLAST (0.466)


B
True Positives
False Positives
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
0
100
200
300
400
COMA (0.318)
HHsearch (0.772)
HHsearch w/o ss (0.529)
COMPASS (0.287)
PSI-BLAST (0.198)

HMM-HMM alignment

Η κατηγορία περιλαμβάνει τις πιο εξελιγμένες μεθόδους που χρησιμοποιούν ειδικούς αλγόριθμους για να στοιχίσουν ολόκληρα HMM

- COACH (Comparison of Alignments by Constructing Hidden Markov Models)
- PRC/webPRC (Profile Comparer)
- HHsearch
- AlignHUSH

COACH

Το COACH είναι μια από τις πρώτες προσπάθειες για στοίχιση profile-profile μέσω των HMM. Τα αρχικά προέρχονται από το Comparison of Alignments by Constructing Hidden Markov Models. Το COACH επιτυγχάνει τη στοίχιση δύο πολλαπλών στοιχίσεων με το να δημιουργεί ένα profile HMM από τη μία εκ των δύο και να στοιχίζει όλες τις αλληλουχίες της άλλης στοίχισης έναντι αυτού του HMM. Η στοίχιση αυτή όμως, δεν γίνεται για κάθε αλληλουχία ξεχωριστά αλλά για όλη τη στοίχιση μαζί, αυτή είναι η βασική διαφορά από τα απλά HMM.

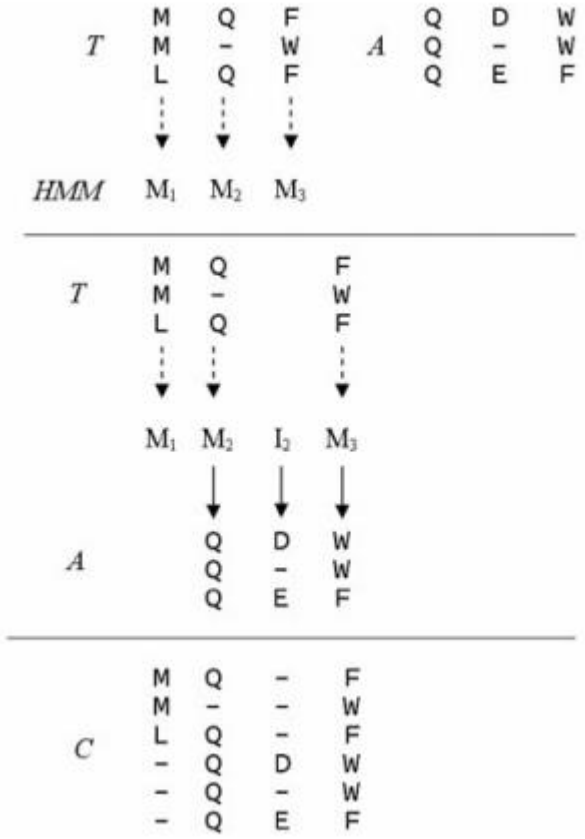
Το COACH όταν συγκρίνεται έναντι των μεθόδων της προηγούμενης γενιάς, δηλαδή του PROF_SIM και του COMPASS με δεδομένα της FSSP database, είναι ικανό κατά μέσο όρο να κατασκευάζει καλύτερες στοιχίσεις με μικρότερα ποσοστά σφάλματος.

www.drive5.com/lobster

Edgar RC, Sjölander K. COACH: profile-profile alignment of protein families using hidden Markov models. Bioinformatics. 2004 May 22;20(8):1309-18

Μέθοδος

- Θεωρεί 2 HMM, το Template (T) και το Input (A), και τα αντιμετωπίζει διαφορετικά. Αφού υπολογιστούν οι παράμετροι του T, και του A, τότε υπολογίζεται η πιθανότητα η στοίχιση A να παράγεται από το T, P (A | HMM). Αυτό είναι διαφορετικό από το να στοιχιθούν οι ακολουθίες μία-μία, καθώς τα αμινοξέα σε μία στήλη θα πρέπει να παράγονται όλα από την ίδια κατάσταση.
- Η τελική στοίχιση γίνεται με την αλγόριθμο Viterbi, αλλά απαιτούνται αλγοριθμικές τροποποιήσεις, με επιπλέον πίνακες δυναμικού προγραμματισμού για να γίνει αυτή η υλοποίηση



webPRC

Το Profile Comparer (PRC) είναι ένα εργαλείο που συγκρίνει δύο HMMs που περιγράφουν πρωτεϊνικές οικογένειες. Το μπορεί να διαβάσει μοντέλα που δημιουργήθηκαν από το SAM και το HMMER, καθώς και checkpoint files (ενδιάμεσα αρχεία) από το PSI-BLAST. Το PRC χρησιμοποιείται από τις βάσεις δεδομένων CATH και Pfam. Καθώς το PRC είναι ένα εργαλείο σύγκρισης προφίλ, αναφέρει μόνο τις στοιχίσεις του προφίλ HMM και δεν παράγει πολλαπλές στοιχίσεις ακολουθιών. Επίσης, είναι διαθέσιμο και σε διαδικτυακή εφαρμογή, ως webPRC server, ο οποίος καθιστά εύκολη την αναζήτηση απομακρυσμένων ομολόγων ή παρόμοιων στοιχίσεων σε διάφορες βάσεις δεδομένων. Επιπλέον, παρέχει τα αποτελέσματα τόσο ως πολλαπλές στοιχίσεις ακολουθιών όσο και ως στοιχισμένα HMM. Επιπλέον, ο χρήστης μπορεί να δει τον σχολιασμό των περιοχών, να αξιολογήσει τα αποτελέσματα του PRC με τον επεξεργαστή πολλαπλών στοιχίσεων Jalview και να δημιουργήσει λογότυπα από τα στοιχισμένα HMM ή τις στοιχισμένες πολλαπλές στοιχίσεις. Έτσι, ο διακομιστής αυτός βοηθά στην ανίχνευση απομακρυσμένων ομολόγων καθώς και στην αξιολόγηση και τη χρήση των αποτελεσμάτων. Η διεπαφή webPRC είναι διαθέσιμη στη διεύθυνση http://www.ibi.vu.nl/programs/prcwww/.

Madera M. Profile Comparer: a program for scoring and aligning profile hidden Markov models. Bioinformatics. 2008 Nov 15;24(22):2630-1. doi: 10.1093/bioinformatics/btn504.

Brandt BW, Heringa J. webPRC: the Profile Comparer for alignment-based searching of public domain databases. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W48-52. doi: 10.1093/nar/gkp279.

Μέθοδος

- Υπολογίζει 3 σκορ, $S_{co\text{-}em}$, S_{simple}, S_{rev}.

$$S_{co\text{-}em}(1,2) = \log \sum_{\sigma} \frac{P(\sigma|\mathrm{HMM1})P(\sigma|\mathrm{HMM2})}{P(\sigma|\mathrm{null})}.$$

- Το S_{simple}, είναι όμοιο αλλά υπολογίζεται μόνο σε περιοχές σημαντικής ομοιότητας, ενώ το S_{rev}:

$$S_{rev}(1,2) = S_{simple}(1,2) - S_{simple}(rev\,1,2),$$

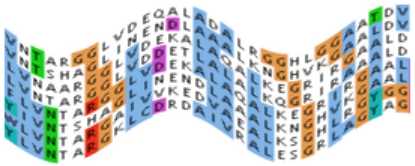
where the reverse HMM is defined as follows:

for every σ, $P(\sigma|rev\mathrm{HMM}) = P(rev\,\sigma|\mathrm{HMM})$.

- Τελικά, η στοίχιση HMM-HMM γίνεται με τον αλγόριθμο Viterbi

webPRC: The Profile Comparer with domain databases

Mail me if you are interested in this server with up-to-date CDD and Pfam models.



The Profile Comparer (PRC, Martin Madera) is a program for aligning and scoring profile hidden Markov models. Here, you can input your query alignment and run PRC against profile HMM libraries of domain databases. Your input alignment can be in (aligned) FASTA, ClustalW, Stockholm, SELEX or GCG MSF format. Using webPRC you can identify and evaluate similar alignments in Pfam, CATH, NCBI's Conserved Domain Database (CDD) TIGRFAMs and SUPERFAMILY, both in HMM and alignment space. Additionally, you can produce logos, and view or download the multiple sequence alignments corresponding to the hits.

Documentation: introduction to the PRC web server, including example output. You can also rerun the example.

References:

Brandt B.W. and Heringa J. (2009) webPRC: webPRC: the Profile Comparer for alignment-based searching of public domain databases. Nucleic Acids Research 37:W48-W52.

Madera M. PRC, the Profile Comparer (we use version 1.5.5 August 2008).

Madera M. (2008) Profile Comparer: a program for scoring and aligning profile hidden Markov models. Bioinformatics 24:2630-2631.

Visualization:
Databases:

Paste in your alignment: example

or upload your alignment:

Browse... No file selected.

Database:

Pfam 25.0 - Mar 2011 (12273 profiles)

PSI-BLAST options: Iterations: 0

E-value: 0.0005

Filters:

PRC options: E-value: 2

Algorithm: forward

Match-match scoring: dot2

Mode: local-local

Other options: Make logos: No Yes

Use --hand: No Yes

Number of hits in graphic: 25

E-mail address (optional):

Run PRC Reset

PRC hit table

Download: PRC scores file and PRC alignments file.
PSI-BLAST: results and generated alignment.

Results of search against Pfam 23.0 - July 2008 (10340 profiles):

hit (hmm2)	description	co-enis	simple	reverse	E-value
VHS	VHS domain: Domain present in VPS-27, Hrs and STAM.	115.9	115.5	99.3	4.7e-44
GAT	GAT domain: The GAT domain is responsible for bindi... >>	74.9	74.7	56.9	8.4e-24
ENTH	ENTH domain: The ENTH (Epsin N-terminal homology) d... >>	30.0	28.9	16.2	0.00022
ANTH	ANTH domain: AP180 is an endocytotic accessory prot... >>	25.3	23.4	13.0	0.0072

Figure 1. An example of the webPRC domain graphic and hit table section for GGA1_HUMAN run against Pfam (after running PSI-BLAST). The graph can be viewed in HMM or alignment space and the hits are hyperlinked to the alignments. The PRC hit table provides links to the original PRC and PSI-BLAST output and shows a table with annotated hits, including the name and, after clicking on '>>', the description from the domain database. The hits are hyperlinked to the source database and *E*-values are hyperlinked to the alignments. Co-emission, simple and reverse scores are calculated by PRC [cf. (12)]. The *E*-value is calculated from the reverse score.

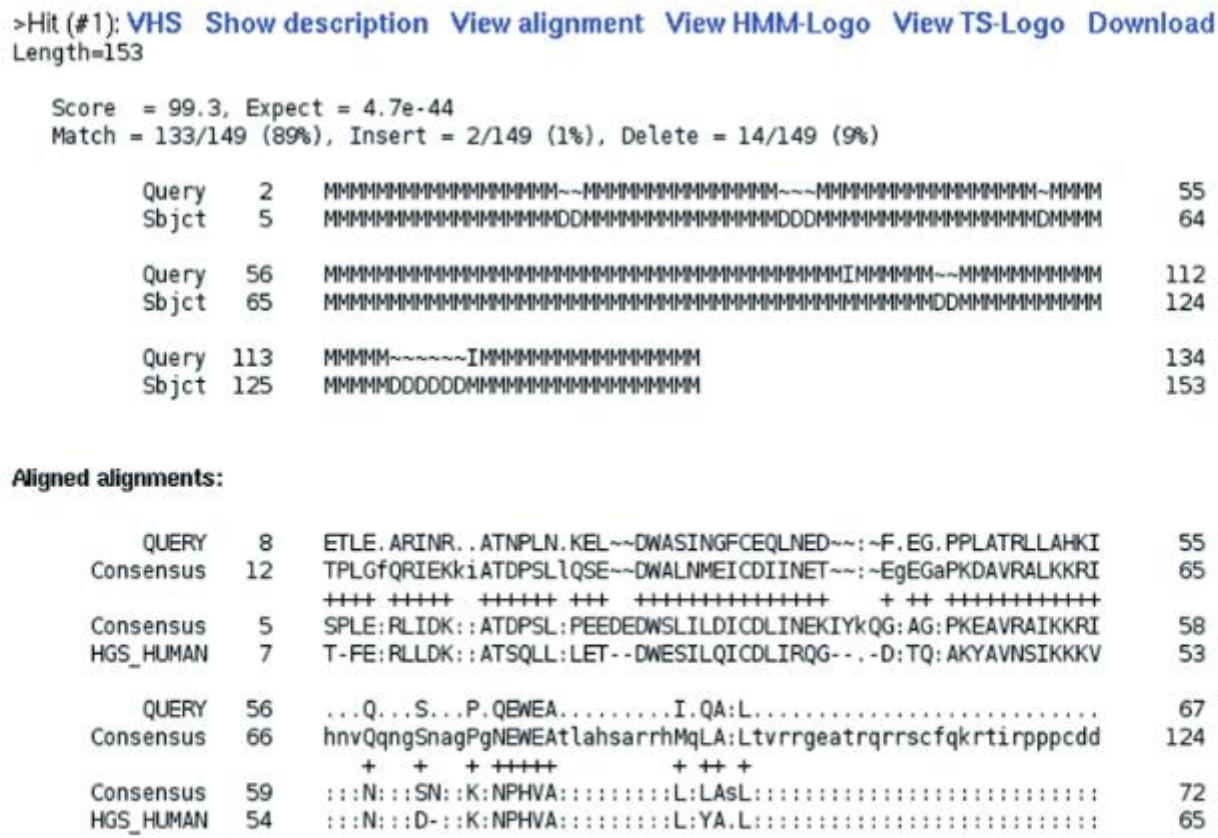


Figure 2. An example alignment showing hit number (#1), links, PRC alignment and aligned alignments (truncated). The original PRC HMM alignment is formatted in a BLAST-like style and now includes the counts and percentages of the Match, Insert and Delete states (M–M, M–I, D–∼ pairs, respectively). The aligned alignments view shows the PRC result in multiple sequence alignment space and includes the first sequence of the query and hit alignment as well as their consensus sequences. The alignments are separated by a mid-line that indicates the PRC match states (M) with a '+'. Gaps present in the seed alignments are indicated by '–', gaps introduced by PRC by '∼' and positions corresponding to columns missing from the HMM by ':'. The entire (aligned) alignments can be viewed with Jalview or downloaded by clicking on 'View alignment' or 'Download', respectively.

HHsearch

Το HHsearch ήταν η πιο σημαντική μέθοδος της πρώτης δεκαετίας του 2000 λόγω του ότι γενίκευσε με εύκολο και αποδοτικό τρόπο την στοίχιση HMM-HMM. Η μέθοδος θεωρεί 2 μοντέλα, το q και το p και υπολογίζει την πιθανότητα μια αλληλουχία x να παράγεται ταυτόχρονα και από τα 2. Χρησιμοποιεί το “log-sum-of-odds score” και περιέχει τις κατάλληλες τροποποιήσεις στους αλγορίθμους δυναμικού προγραμματισμού έτσι ώστε να επιτυγχάνεται αποδοτικά ο υπολογισμός των τυπικών παραμέτρων σε ένα HMM.

Το HHsearch αξιολογήθηκε απέναντι στο BLAST, το PSIBLAST, το HMMER και τις profile–profile μεθόδους PROF_SIM και COMPASS, σε μια σύγκριση 3691 πρωτεϊνικών περιοχών από τη βάση SCOP 1.63 με ομοιότητα μικρότερη του20% και έδωσε εντυπωσιακά αποτελέσματα. Μέχρι σήμερα θεωρείται ίσως η καλύτερη μέθοδος αυτού του είδους.

Söding, J. (2005) Protein homology detection by HMM–HMM comparison Bioinformatics 21, 951-960.

HHsearch

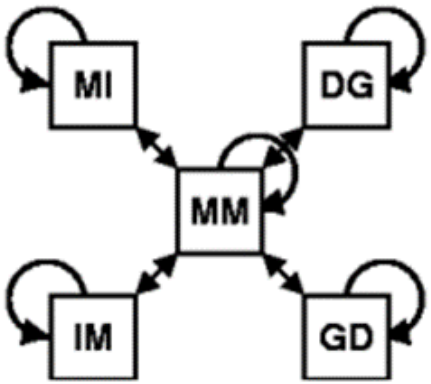
Η πρωτότυπη ιδέα στην οποία βασίστηκε το HHsearch ήταν ότι επιχείρησε να χρησιμοποιήσει αποδοτικά τη στοίχιση HMM-HMM. Αρχικά έδωσαν τον μαθηματικό τύπο για τον υπολογισμό του log-odd scores που απαιτούνται από τον αλγόριθμο Viterbi. Το score από τη στοίχιση 2 στηλών 2 διαφορετικών HMMs (q, t) είναι:

$$S_{aa}(q_i, t_j) = \log_2 \sum_{a=1}^{20} \frac{q_i(a) t_j(a)}{f(a)}$$

Με αυτό το score, η καλύτερη στοίχιση 2 HMMs δίνεται από τη μεγιστοποίηση του log-sum-odds score:

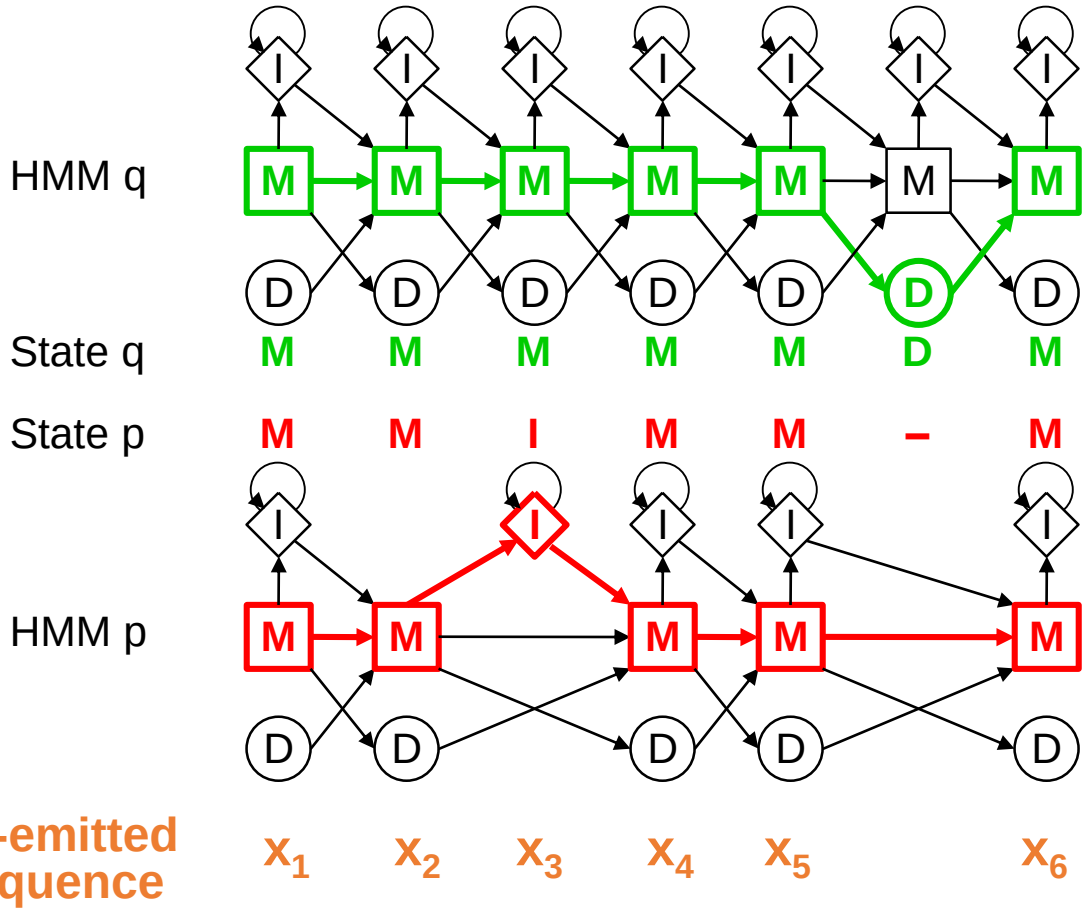
$$S_{\mathrm{LSO}} = \log \sum_{x_1,\ldots,x_L} \frac{P(x_1,\ldots,x_L|\text{co-emission on path})}{P(x_1,\ldots,x_L|\text{Null})}.$$

$$S_{LSO} = \sum_{k:X_k Y_k = MM} S_{aa}(q_{i(k)}, t_{j(k)}) + \log P_{tr}$$



Με αυτό το σύστημα ο αλγόριθμος Viterbi μπορεί να χρησιμοποιηθεί. Για λόγους απλότητας το HHsearch 5 καταστάσεις ταύτισης με τις αντίστοιχες πιθανότητες μετάβασης, έτσι ο δυναμικός προγραμματισμός είχε 5 διαφορετικούς πίνακες.

Find path through two HMMs that maximizes co-emission probability



Include Null model ⇨ maximize "log-sum-of-odds score"

Söding, J. (2005) Protein homology detection by HMM–HMM comparison Bioinformatics 21, 951-960.

Επιπλέον

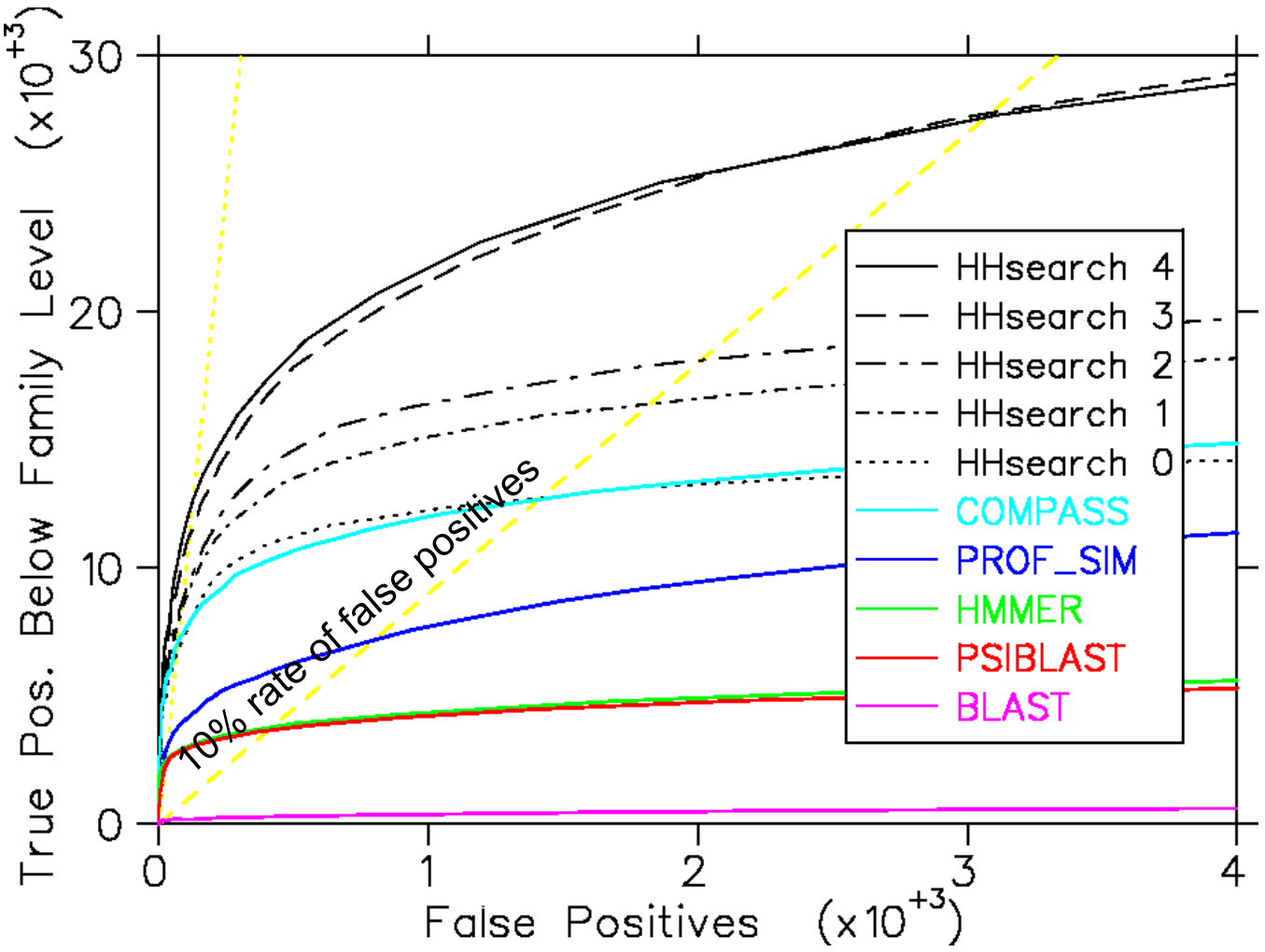
Δύο άλλοι ήταν οι κύριοι παράγοντες που επηρεάζουν το σκορ του HHsearch:

1) σκορ συσχέτισης: βασίζεται στην ιδέα ότι εάν δύο πρωτεΐνες είναι ομόλογες τότε όταν στοιχιθούν οι στήλες με μεγάλο σκορ (συντήρηση) θα πρέπει να συγκεντρώνονται μαζί. Αυτό σημαίνει ότι όσο υψηλότερο αυτό το σκορ, τόσο πιο όμοιες είναι οι αλληλουχίες. Αυτό το σκορ μπορεί απλά να προστεθεί στο «τελικό σκορ καλύτερης στοίχισης» από τον αλγόριθμο Viterbi.

2) Κατά την στοίχιση των δύο στηλών των δύο HMMs, τα στοιχεία δευτεροταγούς δομής (SS) βαθμολογούνται χρησιμοποιώντας σκορ τα οποία λαμβάνουν υπόψη τις τιμές εμπιστοσύνης των προβλέψεων δευτεροταγούς δομής. HHsearch παρέχει δύο κατηγορίες σκορ για σύγκριση δευτεροταγούς δομής: α)- προγνώσεις έναντι προγνώσεων, και β) προγνώσεις έναντι γνωστής δομής. Το δεύτερο χρησιμοποιείται κυρίως κατά την πρόβλεψη της τρισδιάστατης δομής. Αυτές οι προσθήκες προστίθενται στο σκορ του αλγόριθμου Viterbi.

HMM-HMM comparison improves upon profile-profile comparison

All-against-all benchmark on SCOP (20% seq. id.)



HMM-HMM+predSS
HMM-HMM+SS

HMM-HMM+corr

HMM-HMM

profile-profile

profile-profile

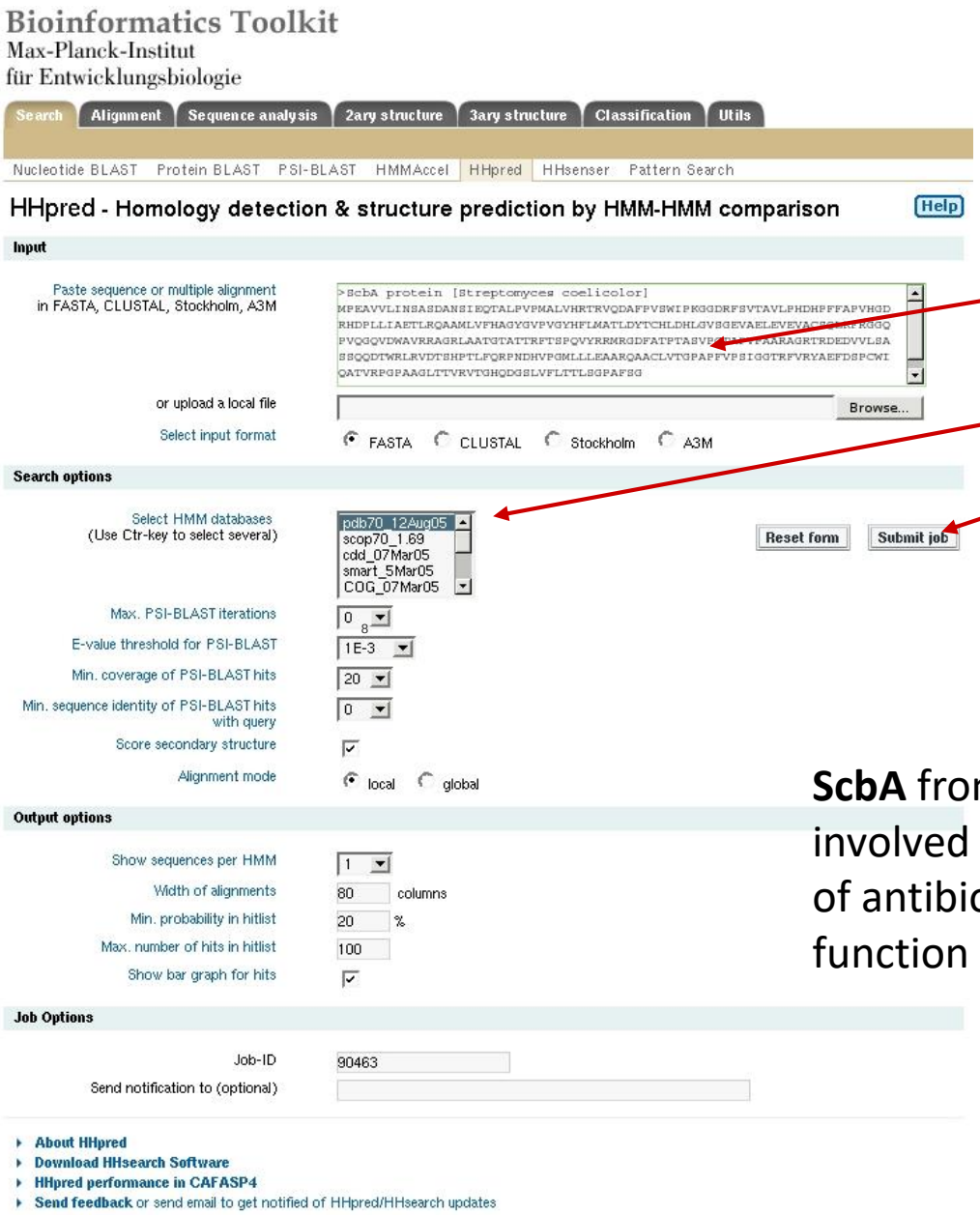
profile-profile

HMM-seq

profile-seq

seq-seq

The HHpred input page



All input parameters are linked to explanations on help pages

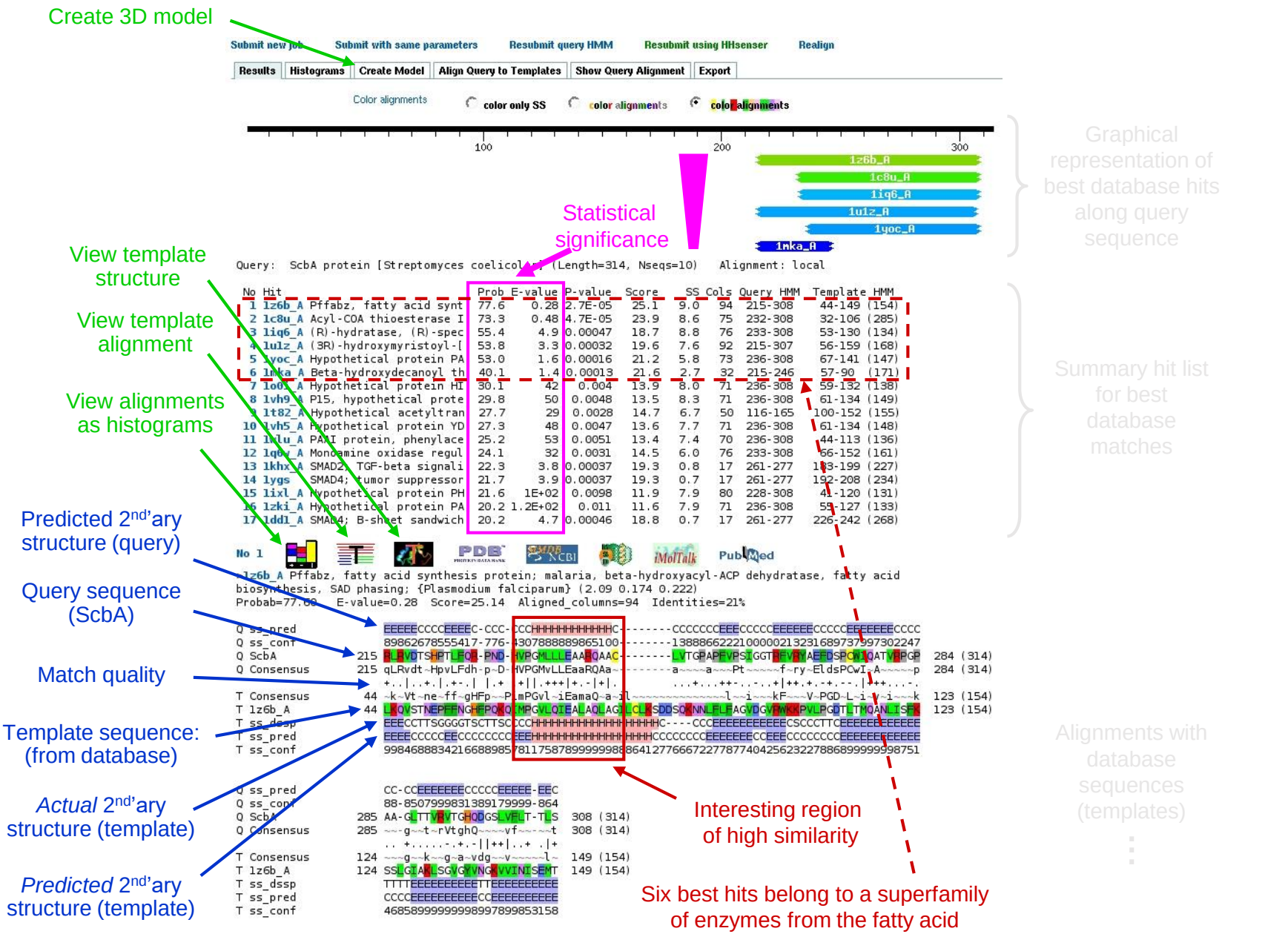
1. Paste ScbA sequence

2. Select database

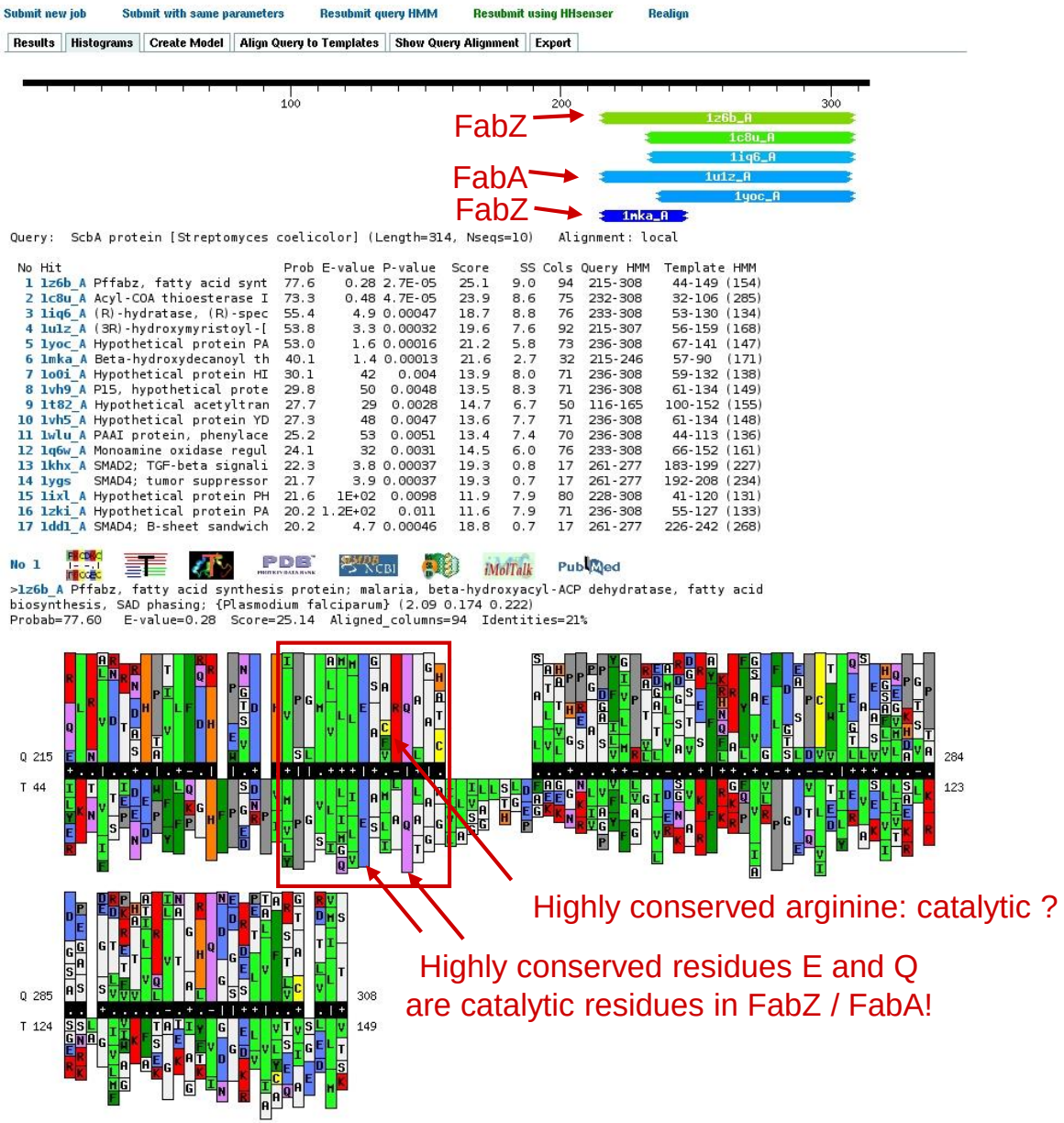
3. Submit job

ScbA from *Steptomyces* is involved in regulating the onset of antibiotics production, but its function is unknown

Search results: alignment view



Histogram view



AlignHUSH

Το AlignHUSH είναι ένα πρόσφατο εργαλείο στοίχισης HMM-HMM το οποίοείναι ικανός να αναγνωρίζει απομακρυσμένες ομοιότητες. Η μέθοδος χρησιμοποιεί δομικές πληροφορίες, υπό τη μορφή προγνώσεων δευτεροταγούς δομής και την υδροφοβικότητα των αμινοξέων για στοίχιση των HMMs που προέρχονται από δύο διαφορετικές πολλαπλές στοιχίσεις . Η επίδραση της χρήσης πληροφοριών των παρακείμενων στήλης της στοίχισης έχει επίσης διερευνηθεί και ευρέθη ότι αυξάνει την ευαισθησία των ευθυγραμμίσεων HMM-HMM και ανίχνευσης απομακρυσμένης ομολογίας. Σε συγκρίσεις στη βάση δεδομένων SCOP το AlignHUSH βρέθηκε να αποδίδει καλύτερα από τις καλύτερες μεθόδους στοίχισης HMM-HMM και παρατηρείται ακόμα μεγαλύτερη ευαισθησία σε υψηλότερα ποσοστά σφάλματος. Η ακρίβεια των στοιχίσεων που έχουν ληφθεί χρησιμοποιώντας το AlignHUSH έχει αξιολογηθεί χρησιμοποιώντας τις δομικές στοιχίσεις που είναι διαθέσιμες στο BaliBASE. Το μήκος στοίχισης και η ποιότητα αυτής βρέθηκαν κατάλληλα για τη μοντελοποίηση με βάση την ομολογία και για σχολιασμό της λειτουργίας πρωτεϊνών. Η ακρίβεια στοίχισης βρέθηκε ότι είναι συγκρίσιμη με τις υπάρχουσες μεθόδους για ευθυγράμμιση προφίλ-προφίλ.

AlignHUSH

Στη μέθοδο AlignHUSH, η ταύτιση μεταξύ των στηλών αποτελείται από τρία διιαφορετικά σκορ, το υδρόφοβο σκορ, το σκορ συντήρησης και το σκορ της δευτεροταγούς δομής. Το σκορ συντήρησης βασίζεται στην εργασία του Soding (HHSearch) και είναι ένα log-sum of odds score. Τα πρότυπα συντήρησης συνήθως παρατηρούνται σε σύντομα μοτίβα και όχι σε απομόνωση και ως εκ τούτου στο AlignHUSH, το σκορ συντήρησης λαμβάνεται σε ένα παράθυρο μερικών στηλών. Έτσι, το σκορ συντήρησης στη θέση (i, j) είναι το άθροισμα των σκορ συντήρησης σε ένα παράθυρο. Η βελτιστοποίηση έχοντας την ευαισθησία ως κριτήριο αποκάλυψε ότι ένα μέγεθος παραθύρου 5 δίνει τα καλύτερα αποτελέσματα και αυτό έχει χρησιμοποιηθεί σε όλη την ανάλυση. Το σκορ υδροφοβικότητας και το σκορ της δευτεροταγούς δομής παράγονται με παρόμοιο τρόπο και το μέγεθος παραθύρου για κάθε σκορ προσδιορίζεται χωριστά.

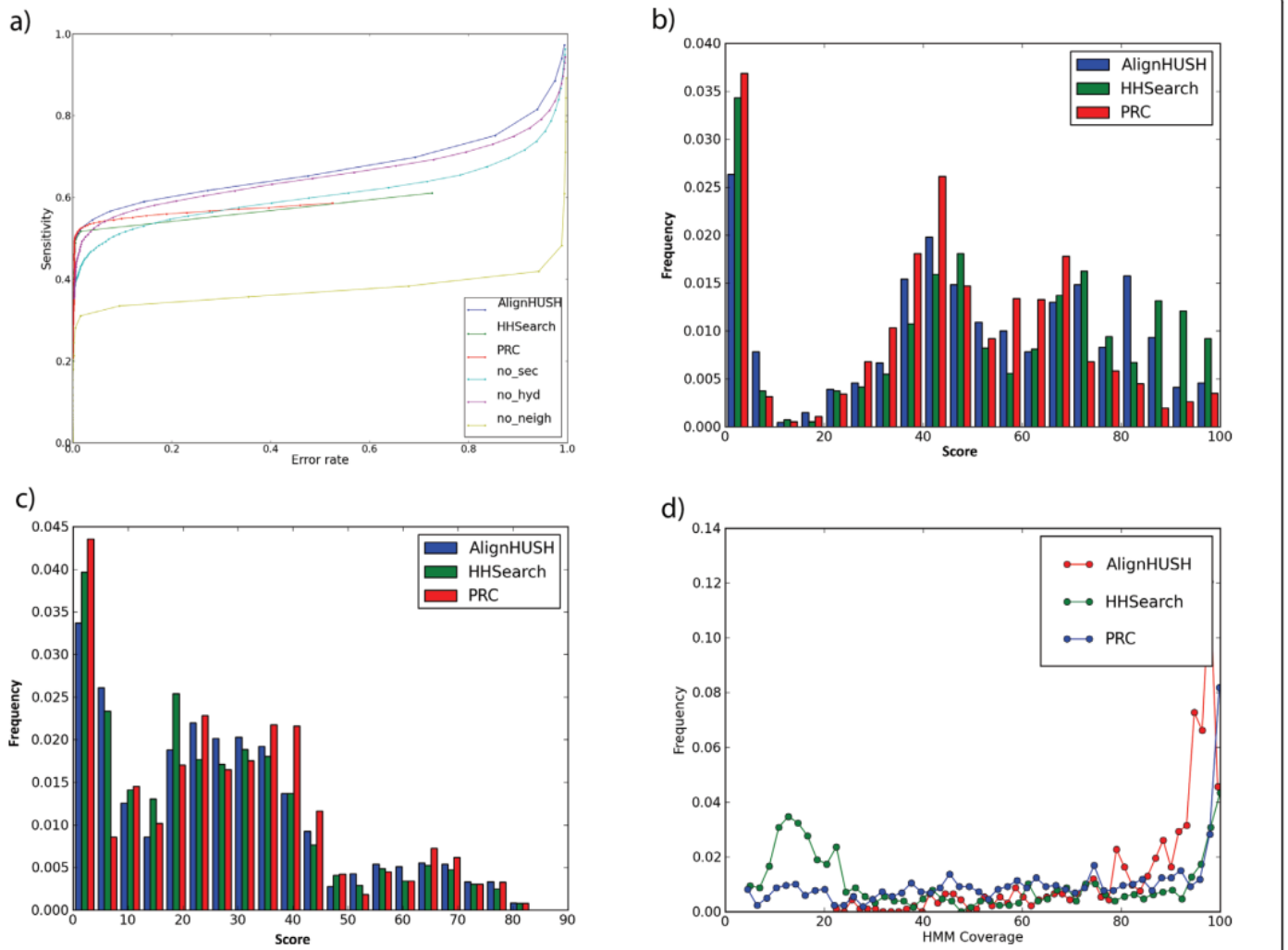


Figure 1 Comparison of performance of AlignHUSH method to HHSearch and PRC. **A)** The sensitivity and error rate values for both AlignHUSH and HHSearch are plotted in this figure. The sensitivity of AlignHUSH is better than HHSearch or PRC at almost all error rates. The 'no_sec', 'no_hyd' and 'no_neigh' are variants of AlignHUSH procedure without use of secondary structure, hydrophobic and neighboring column information respectively. **B)** Alignment accuracy of the three methods that have been examined in detail in the main text. The alignment accuracy given in this plot corresponds to the 'developer score' defined in the main text. The three methods are comparable as far as the accuracy using developer score is concerned. **C)** The alignment accuracy of the three methods using the 'modeller score' defined in the main text. The performance of AlignHUSH is slightly better than that of HHSearch and PRC. HHSearch generated alignments tend to be very short and hence HHSearch has a low value for 'modeller score' alignment accuracy. **D)** The length of the query HMM covered by the alignment is plotted for the alignment between homologous families (two SCOP families belonging to the same SCOP superfamily). The coverage of query HMM is greater in case of AlignHUSH than HHSearch which indicates that AlignHUSH generated alignments are more informative for function annotation, since they cover almost the entire homologous region. The alignment length coverage is very similar between the PRC generated alignments and AlignHUSH generated alignments.

Table 1 Sensitivity at three levels of error-rate for a few profile-profile search methods.

Method	Sensitivity at 5% error rate	Sensitivity at 10% error rate	Sensitivity at 50% error rate	Source
AlignHUSH	52.5%	57.2% (58.5%*)	66.8%	Current work
HHSearch (with SS)	51%	51.3% (48.8%[§], 54.2%*)	56%	Current work and Soding, 2005 [18]
HHSearch (no SS)	NA	46.7%	NA	Soding, 2005 [18]
PRC	53.2%	54.4% (53.2%*)	58.6%	Current work
PROCAIN	NA	52% [#]	~60% [#]	Wang et al, 2009 [25]
PROF_SIM	NA	24.9%	NA	Soding, 2005 [18]
COMPASS	NA	34.0%	NA	Soding, 2005 [18]

The sensitivity values at three levels of error rates for some of the most commonly used profile-profile search methods. The source of each value is given in the last column. Note that the datasets used and the definitions used for true positives can differ from one source to another and hence the values are not comparable across different sources.

"#": The values given for PROCAIN are extracted from Fig 2b of Wang et al [25] where the definition of true positives and false positives is similar to that used in the current work. "§": Value reported from Soding, 2005 [18]. "*": The values reported with asterisk are for the comparison using the latest SCOP release. For

Προσεγγιστικές λύσεις

Μέθοδοι που προσεγγίζουν το πρόβλημα, συνήθως μετατρέποντας ξανά το προφίλ σε μια «ψευτο»ακολουθία ή συναινετική ακολουθία

- PHOG-BLAST
- Quasi consensus alignment
- ConSequenceS (Consensus sequences mimicking profile-profile alignments)
- HHblits

Quasi-consensus-based comparison of HMM

Η μέθοδος αυτή επιτυγχάνει ευαίσθητη ανίχνευση των απομακρυσμένων σχέσεων μεταξύ οικογενειών πρωτεϊνών και την στοίχιση της δομής-αλληλουχίας μέσω μιας απλής προσέγγισης που κάνει χρήση της σύγκρισης των HMM με βάση τις οιονεί-συναινετικές αλληλουχίες (quasi-consensus sequences). Σε μια συγκριτική αξιολόγηση, η προσέγγιση αυτή αποδεικνύεται ότι εντοπίζει με μεγαλύτερη ευαισθησία μακρινές ομοιότητες, ενώ παράγει και στοιχίσεις καλύτερης ποιότητας σε σχέση με παλαιότερες μεθόδους στοίχισης προφίλ-προφίλ που βασίζονται στον δυναμικό προγραμματισμό. Πέραν αυτών, η μέθοδος είναι επίσης σημαντικά ταχύτερη και επομένως είναι κατάλληλη για ένα διακομιστή διαδικτύου. Η μέθοδος είναι διαθέσιμη στη διεύθυνση http://liao.cis.udel.edu/website/servers/modmod

Kahsay RY, Wang G, Gao G, Liao L, Dunbrack R. Quasi-consensus-based comparison of profile hidden Markov models for protein sequences. Bioinformatics. 2005 May 15;21(10):2287-93.

Quasi-consensus-based comparison of HMM

Ένα πρόβλημα με τη χρήση συναινετικών ακολουθιών (consensus sequences) για τη σύγκριση των προφίλ HMMs είναι ότι το να βρεθεί η ακριβής συναινετική αλληλουχία για ένα HMM είναι ένα πρόβλημα NP-hard. Ως υποκατάστατο της ακριβούς συναινετικής ακολουθίας, με μια οιονεί συναινετική ακολουθία (quasi-consensus sequence). υπολογιζόμενη από μια ευρετική προσέγγιση που χρησιμοποιεί το πρόγραμμα hmmemit από το πακέτο HMMER με την επιλογή -c. Αυτό προκύπτει από την κατανομή συχνοτήτων των αμινοξέων στις καταστάσεις match, insert του HMM. Εάν η κατάσταση ταύτισης (match) σε έναν κόμβο έχει πιθανότητα ≥0,5, τότε αυτή η κατάσταση είναι «συναίνεση» και επιλέγεται το κατάλοιπο με τη μέγιστη πιθανότητα. Από την άλλη πλευρά, αν η κατάσταση insert (εισαγωγής) χρησιμοποιείται με πιθανότητα ≥0,5, τότε η θέση αυτή χαρακτηρίζεται «συναινετική» και χρησιμοποιείται το σύμβολο («X»).

PHOG-BLAST

Το PHOG-BLAST μετατρέπει το προφίλ σε ένα πεπερασμένο αλφάβητο και χρησιμοποιεί το hash για γρήγορη αναζήτηση. Για να προσδιοριστεί το βέλτιστο αλφάβητο, έγιναν αναλύσεις στις στήλες αξιόπιστων πολλαπλών στοιχίσεων και εφαρμόστηκε μια ειδική διαδικασία ομαδοποίησης. Δείχνουμε ότι η διαδικασία ομαδοποίησης λειτουργεί καλύτερα εάν οι παράμετροί της επιλέγονται έτσι ώστε να αποκτούνται 20 ομάδες (clusters) προφίλ που μπορούν να ερμηνευτούν ως προγονικά αμινοξέα. Έγινε σύγκριση του PHOG-BLAST έναντι του PSI-BLAST σε τρεις γνωστές βάσεις δεδομένων πολλαπλών στοιχίσεων: COG, PFAM και BALIBASE. Στη βάση δεδομένων COG και οι δύο αλγόριθμοι είχαν την ίδια απόδοση, στο PFAM και το BALIBASE το PHOG-BLAST ήταν πολύ ανώτερο από το PSI-BLAST. Το PHOG-BLAST απαιτούσε επιπλέον 10-20 φορές λιγότερη μνήμη υπολογιστή και χρόνο υπολογισμού από το PSI-BLAST. Το PHOG-BLAST είναι λιγότερο ακριβές από την αυστηρή μέθοδο σύγκρισης προφίλ-προφίλ που βασίζεται στον δυναμικό προγραμματισμό, αν και τρέχει πολύ πιο γρήγορα.

Merkeev IV, Mironov AA. PHOG-BLAST--a new generation tool for fast similarity search of protein families. BMC Evol Biol. 2006 Jun 22;6:51.

Table 1: This table shows the ability of PHOG-BLAST and PSI-BLAST to match members of different subalignments belonging to one protein family against each other as BBHs when the initial multiple alignment of the protein family was split in two subalignments. See the Results section for explanation

Test index	Database	Total number of BBHs to be found	Number of BBHs found	
			PHOG-BLAST	PSI-BLAST
1	COG	3164	3096	3109
2	PFAM	7315	6773	4278
3	BALIBASE	143	70	0

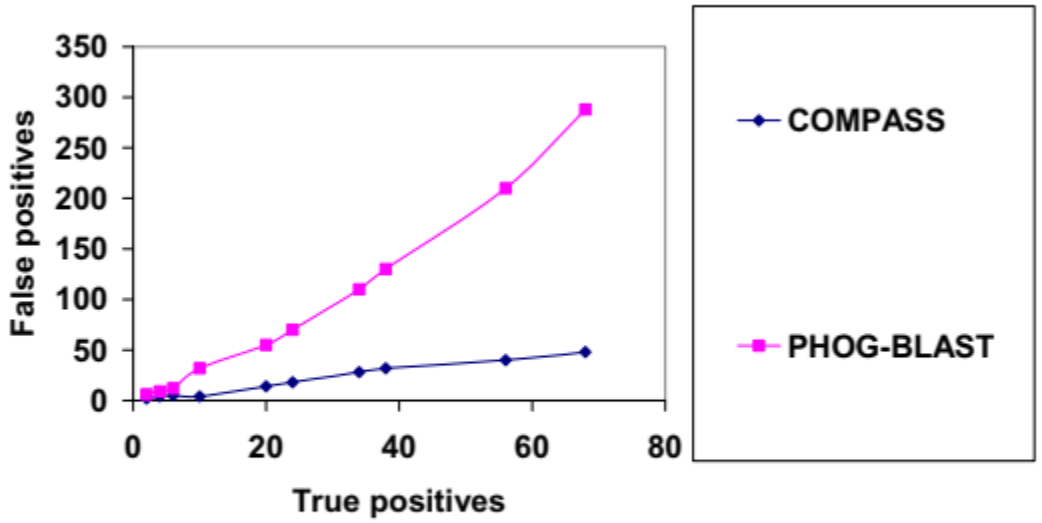


Figure 4
Sensitivity curves of COMPASS and PHOG-BLAST for the remote homology test between PFAM alignments.

mimicking profile-profile alignments

Οι καλύτερες μέθοδοι που προσδιορίζουν την ομοιότητας αλληλουχίας μέσω συγκρίσεων προφίλ-προφίλ είναι πολύ πιο αργές και πιο πολύπλοκες από τις συγκρίσεις αλληλουχίας-αλληλουχίας και προφίλ-αλληλουχίας όπως, αντίστοιχα, το BLAST και το PSI-BLAST. Οι οικογένειες των σχετικών γονιδίων και των γονιδιακών προϊόντων (πρωτεΐνες) μπορούν όμως να εκπροσωπούνται από συναινετικές αλληλουχίες (consensus sequences) που περιέχουν το αμινοξύ που είναι συχνότερο σε κάθε θέση της πολλαπλής στοίχισης στην οικογένεια αυτή. Αυτή η προσέγγιση βελτίωσε τις αναζητήσεις και τις στοιχίσεις ως ένα τυπικό πρόσθετο στο PSI-BLAST χωρίς αλλαγές στον κώδικα. Οι βελτιώσεις ήταν ιδιαίτερα σημαντικές για πιο δύσκολες εργασίες, όπως η αναγνώριση των μακρινών δομικών σχέσεων μεταξύ των πρωτεϊνών και των αντίστοιχων στοιχίσεων. Παρά το γεγονός ότι οι βελτιώσεις ήταν υψηλότερες για πιο αποκλίνουσες σχέσεις, ήταν συνεπείς ακόμη και με υψηλή ακρίβεια / χαμηλό ποσοστό σφαλμάτων για μη συγγενείς πρωτεΐνες. Οι βελτιώσεις ήταν πολύ εύκολο να επιτευχθούν. δεν τροποποιήθηκε καμία παράμετρος που χρησιμοποιήθηκε από το PSI-BLAST και δεν άλλαξε ούτε μία γραμμή κώδικα. Επιπλέον, το πρόσθετο αυτό απαιτεί σχετικά λίγο επιπλέον χρόνο CPU. Οι καλοί χρήστες του PSI-BLAST μπορούν άμεσα να επωφεληθούν από τη χρήση συναινετικών ακολουθιών στους τοπικούς τους υπολογιστές. https://rostlab.org/owiki/index.php/ConSequenceS

Przybylski D, Rost B. Consensus sequences improve PSI-BLAST through mimicking profile-profile alignments. Nucleic Acids Res. 2007;35(7):2238-46.

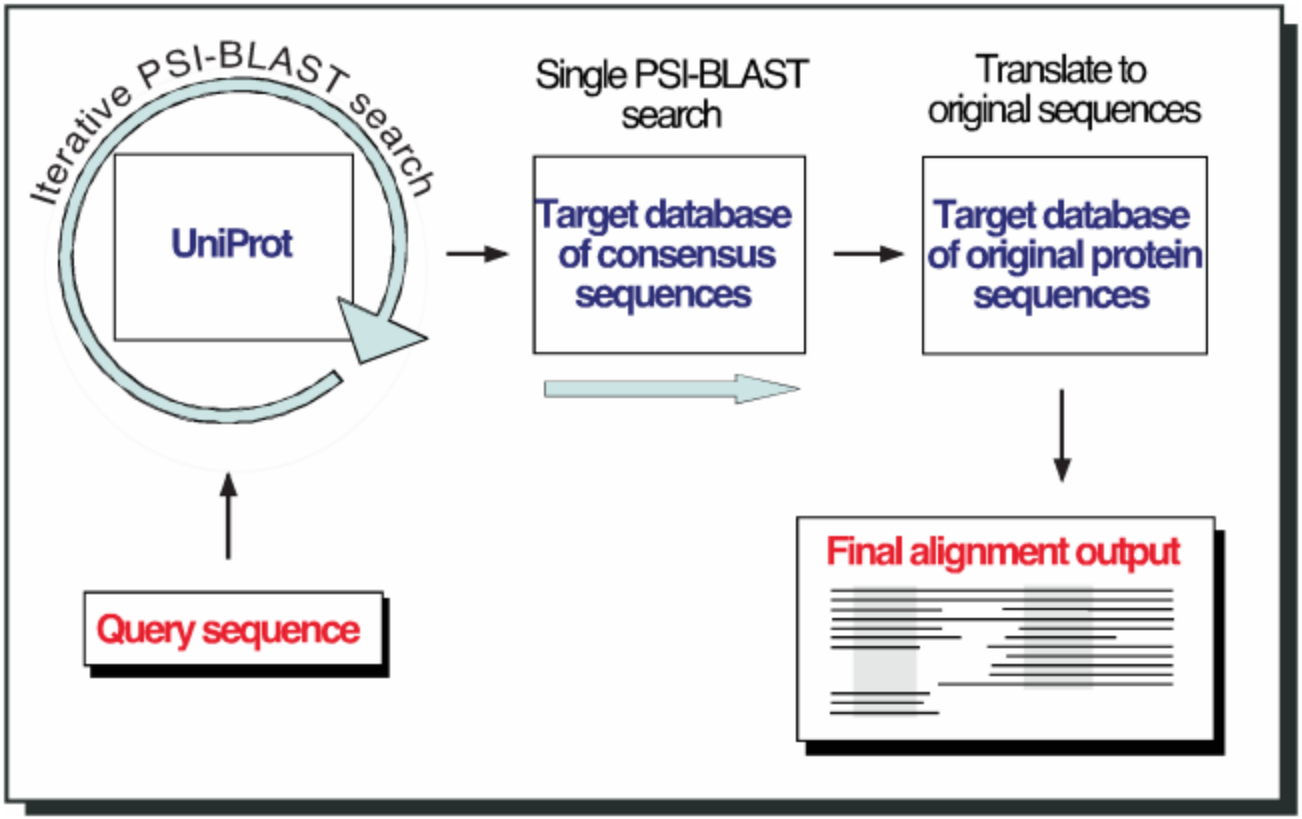
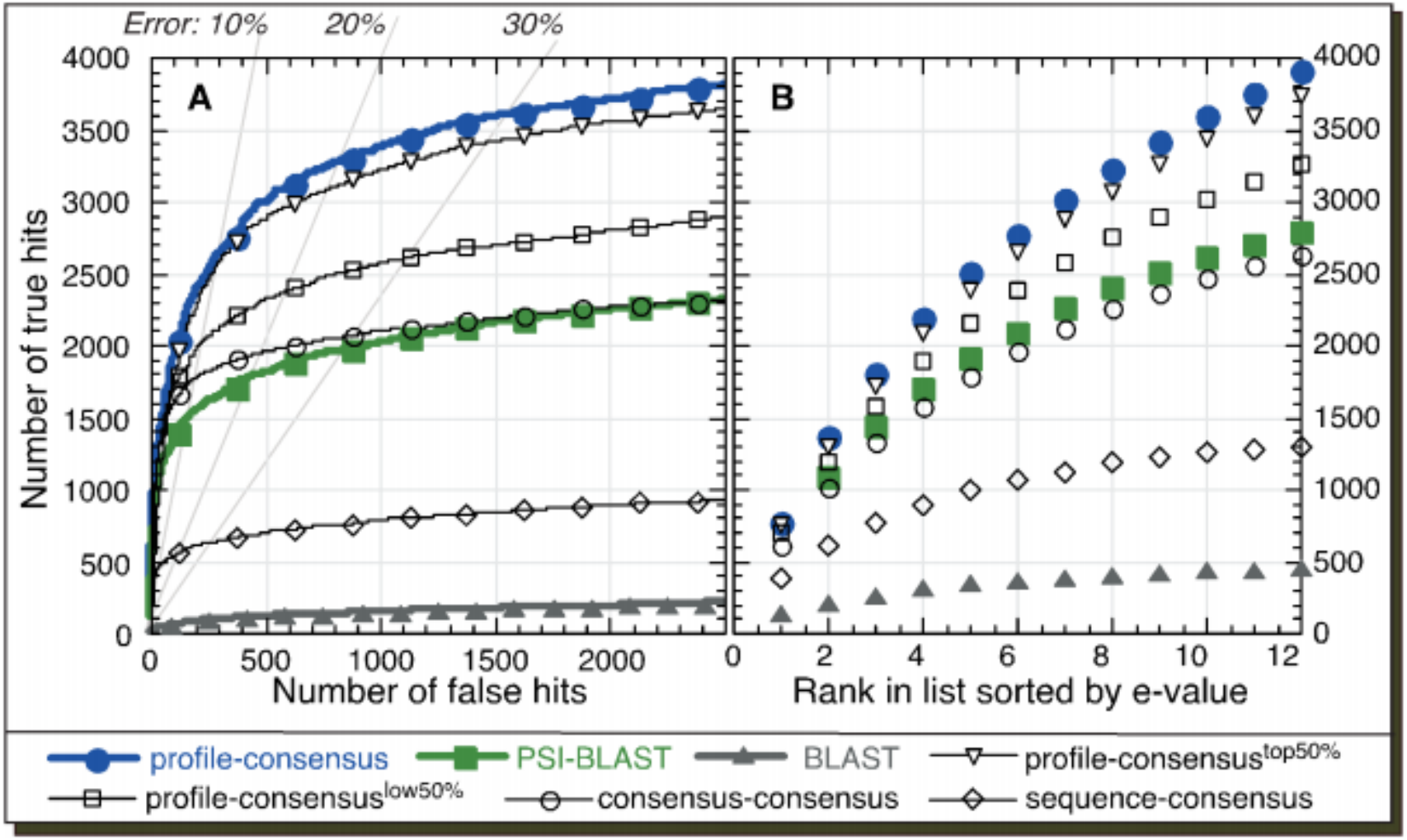


Figure 1. Sketch of consensus search. First, the PSSM for a query protein sequence is built by an iterative PSI-BLAST search over a large database of proteins sequences (such as UniProt). The resulting PSSM is then used to search and align sequences contained in a target database of consensus sequences. Finally, consensus sequence alignments are translated to alignments of the native raw protein sequences.

```
This file contains alignments of the query sequence with consensus sequences.
On the left side:  consensus sequences are translated back into raw (real) sequences.
On the right side: consensus sequences are not translated.
                   *******

-------- Raw sequences  --------

BLASTP 2.2.9 [May-01-2004]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= 1cyx   mol:protein length:205     Cyoa
         (205 letters)

Database: pdbcons_100
           22,314 sequences; 5,192,384 total letters

Searching.........................................done



                                                                    Score    E
Sequences producing significant alignments:                        (bits) Value

1iby_A                                                                100   2e-22


>1iby_A
          Length = 112

 Score =  100 bits (249), Expect = 2e-22
 Identities = 24/87 (27%), Positives = 36/87 (41%), Gaps = 3/87 (3%)

Query: 31  IYPEQGIATVNEIAFPANTPVYFKVT-SNSVMHSFFIPRLGSQIYAMAGMQTRLHLIANE 89
           +   + +   N +  P   PV + +T S+ V H ++IP  G ++ A  GM        N
Sbjct: 28  IRAFNVLNEPETLVVKKGDAVKVVVENKSPISEGFSIDAFGVQEVIKAGETKTISFTADK 87

Query: 90  PGTYDGICAEICGPGHSGMKFKAIATP 116
           P  Y G C+E CG  H  M
Sbjct: 88  AGAFTIWCQLHPKNIH--LPGTLNVVE 112



  Database: pdbcons_100
    Posted date:  Sep 25, 2006  3:27 PM
  Number of letters in database: 5,192,384
  Number of sequences in database:  22,314

Lambda     K      H
   0.322    0.150    0.450

Lambda     K      H
   0.267   0.0460    0.140

-------- Consensus sequences (in the "Sbjct:" fields) --------

BLASTP 2.2.9 [May-01-2004]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= 1cyx   mol:protein length:205     Cyoa
         (205 letters)

Database: pdbcons_100
           22,314 sequences; 5,192,384 total letters

Searching.........................................done



                                                                    Score    E
Sequences producing significant alignments:                        (bits) Value

1iby_A                                                                100   2e-22


>1iby_A
          Length = 112

 Score =  100 bits (249), Expect = 2e-22
 Identities = 24/87 (27%), Positives = 36/87 (41%), Gaps = 3/87 (3%)

Query: 31  IYPEQGIATVNEIAFPANTPVYFKVT-SNSVMHSFFIPRLGSQIYAMAGMQTRLHLIANE 89
           +   + +   N +  P   PV + +T S+ V H ++IP  G ++ A  GM        N
Sbjct: 28  MNQFRLLEVDNRLVVPMGDPVRWVLTNSDDVWHGWWIPSHGIKMDACHGMTWTYWFTFNR 87

Query: 90  PGTYDGICAEICGPGHSGMKFKAIATP 116
           P  Y G C+E CG  H  M
Sbjct: 88  PWMYYGQCSEYCGANH--MPGVVEVVE 112



  Database: pdbcons_100
    Posted date:  Sep 25, 2006  3:27 PM
  Number of letters in database: 5,192,384
  Number of sequences in database:  22,314

Lambda     K      H
   0.322    0.150    0.450

Lambda     K      H
   0.267   0.0460    0.140
```


Error: 10%
20%
30%
A
B
Number of true hits
Number of false hits
Rank in list sorted by e-value
0
500
1000
1500
2000
2500
3000
3500
4000
2
4
6
8
10
12
profile-consensus
PSI-BLAST
BLAST
profile-consensus^top50%
profile-consensus^low50%
consensus-consensus
sequence-consensus

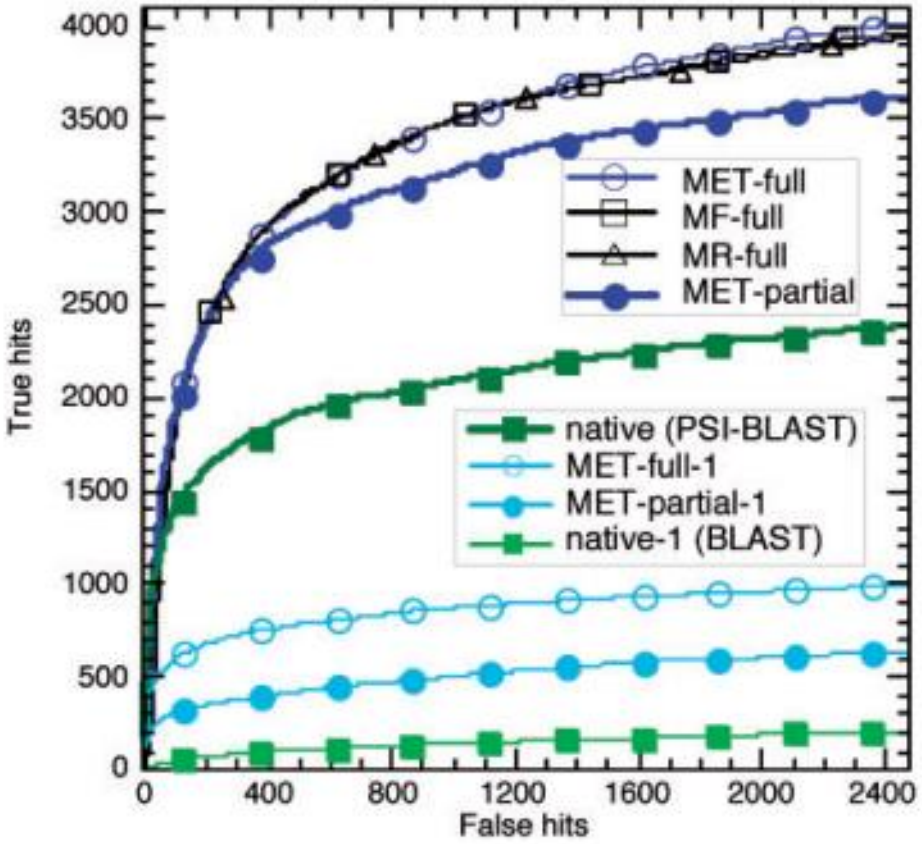


Fig. 2. Comparison of search performance. All-against-all alignments of the test set sequences were ordered by their PSI-BLAST *E*-values. The cumulative numbers of non-trivial true relations (same SCOP superfamily but different SCOP family) were plotted against the cumulative numbers of false positives (different SCOP-folds). The profile-sequence searches against the full consensus sequences performed best (top three curves: *MET-full*, *MF-full*, *MR-full*). Profile-sequence searches against partial consensus sequences were slightly less efficient (*MET-partial*), but they were still significantly better than standard profile-sequence (*native*). Sequence–sequence searches (one cycle of PSI-BLAST with BLOSUM62 matrix) were clearly inferior (*MET-full-1*, *MET-partial-1*, *native-1*).

HHblits

Η μέθοδος στοίχοσης HHM-HMM HHsearch5 χρησιμοποιείται από πολλούς από τους καλύτερους εξυπηρετητές πρόβλεψης δομής πρωτεϊνών, μεταξύ των οποίων και ο HHpred6, οποίος θεωρείται ο κορυφαίος διακομιστής πρόβλεψης δομής πρωτεϊνών μέσω της προτυποποίησης με βάση την ομολογία. Ωστόσο, αυτές οι μέθοδοι είναι γενικά πολύ αργές για επανειλημμένη αναζήτηση σε μεγάλες βάσεις δεδομένων αλληλουχιών όπως η UniProt ή οι βάσεις του NCBI. Η μέθοδος Hhblits (HMM-HMM–based lightningfast iterative sequence search), επεκτείνει το HHsearch με σκοπό να κάνει εφικτές γρήγορες, επαναληπτικές αναζητήσεις σε μεγάλες βάσεις. Το προφίλτρο στοίχισης προφίλ προφίλ HHblits μειώνει τον αριθμό των πλήρων στοιχισμένων HMM-HMM από πολλά εκατομμύρια σε μερικές χιλιάδες, καθιστώντας το ταχύτερο από το PSI-BLAST αλλά παραμένει τόσο ευαίσθητο όσο το Hhsearch.

Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011 Dec 25;9(2):173-5. doi: 10.1038/nmeth.1818.

HHblits

Το HHblits μετατρέπει πρώτα την αλληλουχία επερώτησης (ή την πολλαπλή στοίχιση) σε HMM. Αυτό συμβατικά συμβαίνει με την προσθήκη ψευδοτιμών αμινοξέων που είναι φυσικοχημικά όμοια με το αμινοξύ στην επερώτηση. Το HHblits υπολογίζει τιμές που εξαρτώνται από το τοπικό πλαίσιο αλληλουχίας (δηλαδή τις 13 θέσεις γύρω από κάθε κατάλοιπο). Αυτή η μέθοδος είχε βελτιώσει σημαντικά την ευαισθησία και την ποιότητα στοίχισης του προφίλ. Στη συνέχεια, το HHblits αναζητά στη βάση δεδομένων HMM και προσθέτει τις ακολουθίες από τα HMM με σκορ κάτω από ένα καθορισμένο όριο αναμενόμενης τιμής (τιμή E). Για ταχύτητα και ευαισθησία, το προφίλτρο που χρησιμοποιεί είναι κρίσιμο. Η βασική ιδέα ήταν να εφαρμοστεί η σύγκριση προφίλ-προφίλ ως σύγκριση αλληλουχίας-προφίλ με διακριτοποίηση των διανυσμάτων με τις πιθανότητες των 20 αμινοξέων σε κάθε στήλη HMM, μετατρέποντας το σε ένα αλφάβητο με 219 γράμματα. Κάθε γράμμα αντιπροσωπεύει μια τυπική στήλη του προφίλ. Με αυτόν τον τρόπο η βάση δεδομένων των HMM αντικαθίσταται από αλληλουχίες σε αυτό το εκτεταμένο αλφάβητο, αγνοώντας τις πιθανότητες εισαγωγής και απαλοιφής των HMMs. Πριν από την προ-φιλτράρισμα, ο αλγόριθμος υπολογίζει το σκορ κάθε στήλης του HMM της επερώτησης με κάθε ένα από τα 219 γράμματα, πράγμα που έχει ως αποτέλεσμα ένα εκτεταμένο προφίλ αλληλουχίας 219 σειρών.

$$S_{ik} = \log_2 \sum_{a=1}^{20} q_i(a) p_k(a) / f(a)$$

HHblits

Μία αναζήτηση μίας επανάληψης με το HHblits στην έκδοση 2.2.17 μέσω UniProt20 (2.6 εκατομμύρια ομάδες και 15 εκατομμύρια ακολουθίες) για 100 τυχαία επιλεγμένες αλληλουχίες επερώτησης έλαβε διάμεσο 31 δευτερόλεπτα και μέσο όρο 1 λεπτό 13 δευτερόλεπτα σε έναν μονοπύρηνο Xeon 2.9 GHz. Για μία επανάληψη αναζήτησης μέσω της UniProt (15 εκατομμύρια ακολουθίες), το PSI-BLAST χρειάστηκε 1 λεπτό 7 δευτερόλεπτα (μέσος όρος) και 1 λεπτό 26 δευτερόλεπτα (μέσο όρο) και το HMMER3 χρειάστηκε 2 λεπτά 57 δευτερόλεπτα (μέσος όρος) και 5 λεπτά 8 δευτερόλεπτα . Οι συμπληρωματικές επαναλήψεις έλαβαν περίπου το ίδιο χρονικό διάστημα με την πρώτη επανάληψη και ως εκ τούτου συνολικά τα HHblits ήταν περίπου δύο φορές (15%) τόσο γρήγορα όσο το PSI-BLAST και ήταν 6 φορές(διάμεσος) και 4 φορές ταχύτερα από το HMMER3

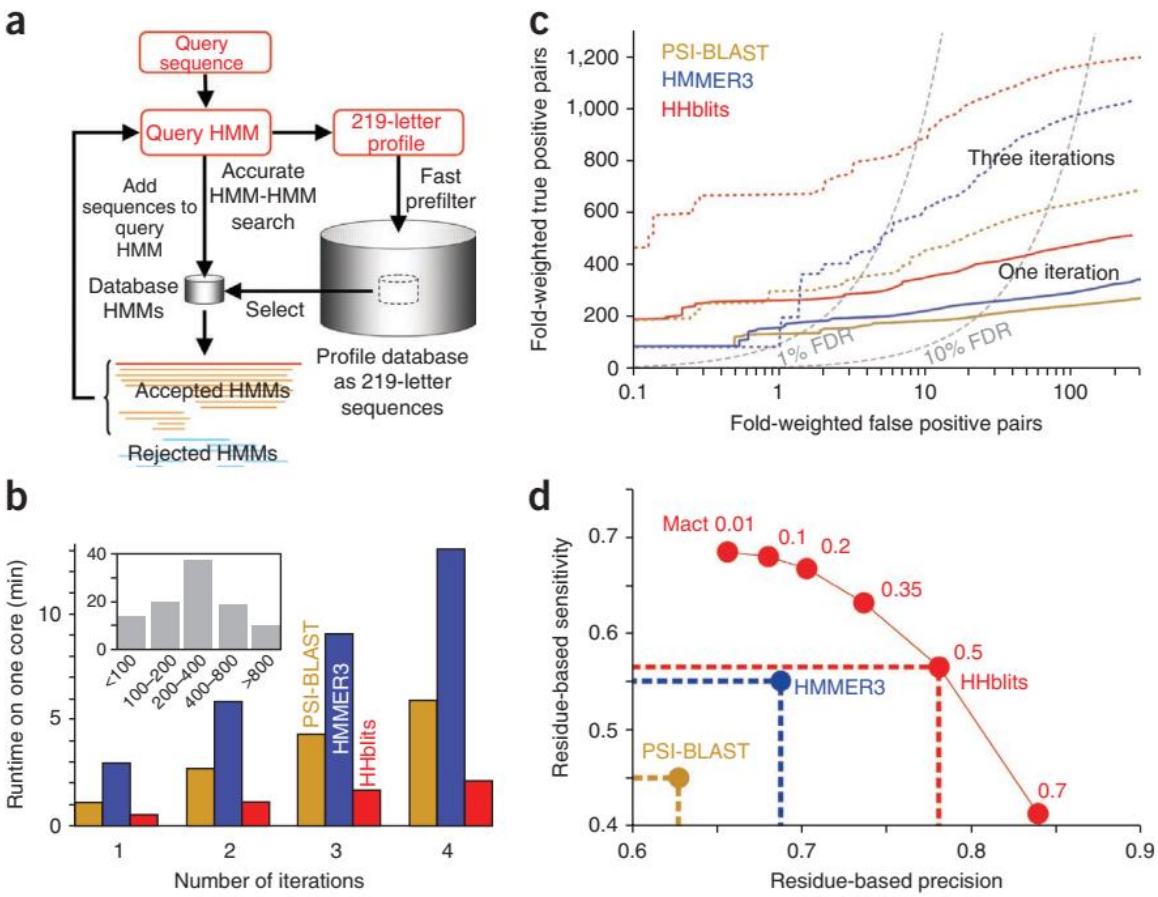


Figure 1 | Workflow and benchmark comparison. (**a**) HHblits can iteratively search for homologous sequences in large databases such as UniProt. The HHblits database is a clustered version in which each set of full-length alignable sequences is represented by an HMM. Sequences from matched HMMs with a statistically significant *E* value are added to the query MSA, from which a new HMM is calculated for the next search iteration. A prefilter reduces the number of full HMM-HMM alignments by ~2,500-fold. (**b**) Median run times for searches with 100 test sequences through the UniProt or UniProt20 database (the inset shows the test sequence length distribution). (**c**) True positive pairs (same SCOP fold) compared to false positive pairs (different SCOP fold) for one and three search iterations in an all-against-all comparison. FDR, false discovery rate. (**d**) Mean fraction of correctly aligned residue pairs out of all structurally alignable pairs (sensitivity) compared to the fraction of correctly aligned pairs out of all the aligned pairs (precision). The parameter mact controls the alignment greediness (**Supplementary Fig. 10**).

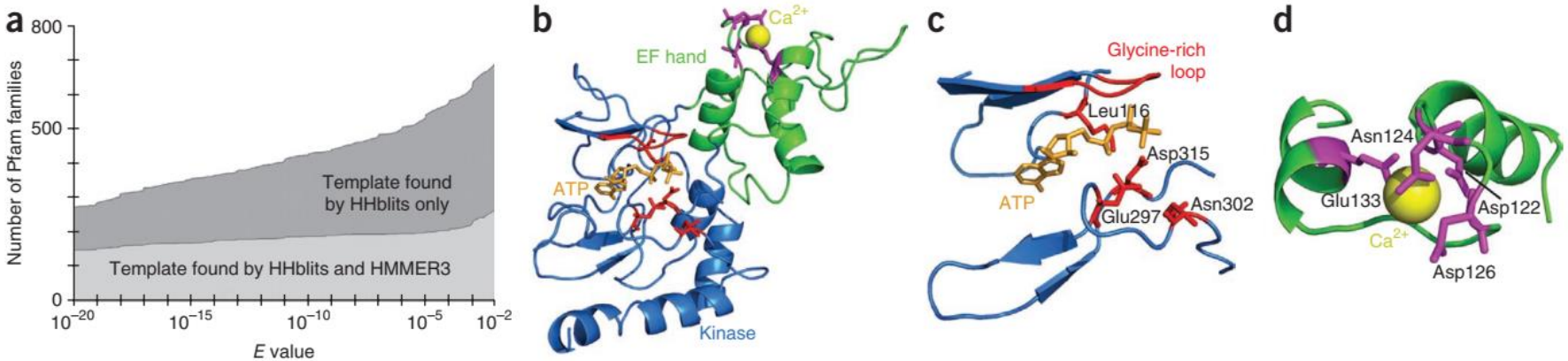


Figure 2 | Structure predictions for Pfam families and the modeling of human Pip49 (also known as FAM69B). (**a**) Families to which only HHblits and both HHblits and HMMER3 assigned a structural template below a given *E* value. (**b**) Homology model of human Pip49 kinase domain (blue) with the inserted EF hand (green). (**c**) Catalytic center showing the conserved residues (red) for protein kinase activity. (**d**) EF hand insertion with the conserved residues (magenta) for the predicted Ca^{2+}-dependent activation.

Οπτικοποίηση HMM-HMM

Η διαθεσιμότητα προχωρημένων εργαλείων σύγκρισης προφίλ-προφίλ, όπως το PRC ή το HHsearch, απαιτεί και εξελιγμένα εργαλεία οπτικοποίησης. Για το σκοπό αυτό χρησιμοποιούνται προσεγγίσεις βασισμένες στην έννοια των λογοτύπων HMM (HMMlogo) επεκτείνοντας την σε ζευγάρια. Η μέθοδος απεικονίζει τις ομοιότητες των ζευγών των προφίλ των οικογενειών πρωτεϊνών με διαισθητικό τρόπο. Δύο λογότυπα HMM, ένα για κάθε προφίλ, σχεδιάζονται το ένα πάνω στο άλλο και στη συνέχεια, οι στοιχισμένες καταστάσεις επισημαίνονται και συνδέονται.

Η μέθοδος είναι διαθέσιμη στη διεύθυνση: http://www.sanger.ac.uk/Software/analysis/logomat-p Υπάρχει όμως διαθέσιμη και τοπική έκδοση

Benjamin Schuster-Böckler, Alex Bateman; Visualizing profile–profile alignment: pairwise HMM logos, Bioinformatics, Volume 21, Issue 12, 15 June 2005, Pages 2912–2913, https://doi.org/10.1093/bioinformatics/bti434

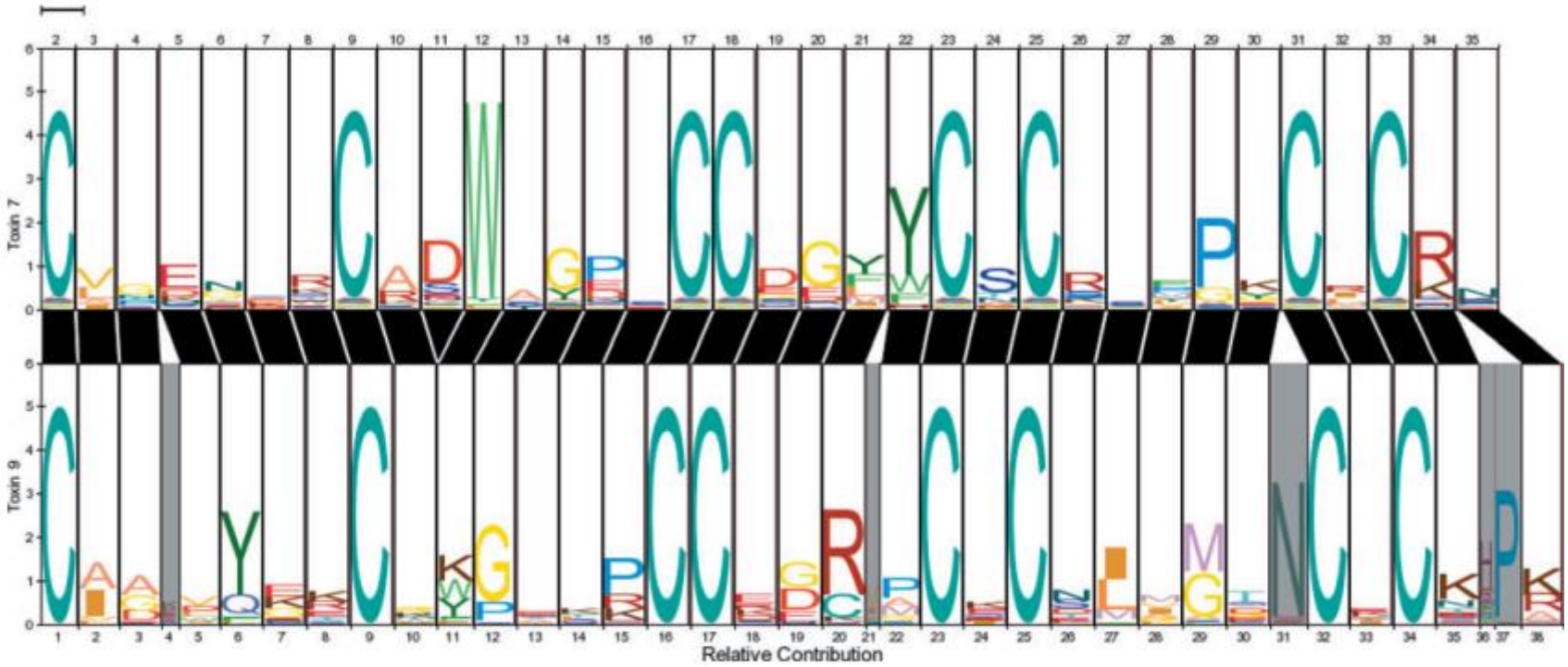


Fig. 1. Alignment of the Toxin_7 against the Toxin_9 Pfam family. For each family, an HMM logo is drawn. The numbers above and below each logo show state positions in the HMM. The overall height of the letter stacks represents the information content, the relative letter height corresponds to its emission probability. The column width denotes the relative contribution, the product of the probability that the state is traversed with the expected number of self transitions for the respective state. This is to account for the varying length of insertions. Insert states are drawn in *red*. Frequently, their relative contribution is very small, making them hard to see. In this picture, you find narrow insert states e.g. at positions 27 and 28 of the Toxin_7 family. The aligned states in each HMM are framed and connected by a block. Omitted states are shaded in *grey*.