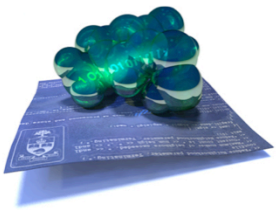
A
Bioinformatics
Course

PSI-BLAST



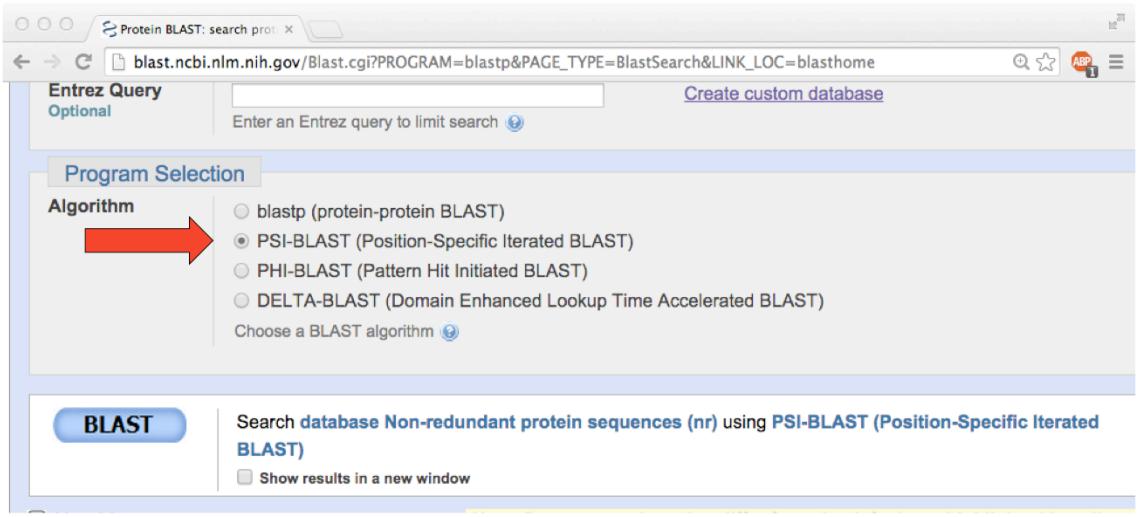
Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

Searching with **profiles** instead of individual sequences is more sensitive!

It reduces the effect of *neutral drift* in individual sequences and focuses the search on the *commonalities* and the specific *tolerance to variation* that is particular to a *family* of homologous sequences (in which all members have diverged from a common ancestor).

Invoking PSI-BLAST is trivial:



Initiate a PSI-BLAST search simply by choosing the option on the BLAST input form.

But note: **invoking** the algorithm is trivial. Using it **correctly** and interpreting the results, perhaps not so much.

PSI-BLAST proceeds in five steps:

1. Select a query and BLAST it against a protein database
2. PSI-BLAST constructs a multiple sequence alignment from the BLAST hits, then creates a "profile" (or position-specific scoring matrix (PSSM))
3. The PSSM is used as a query against the database
4. PSI-BLAST estimates statistical significance (E values) and proposes significant hits for inclusion into the next iteration
5. Repeat steps [3] and [4] iteratively, typically 5 times. At each new search, a new profile is constructed from the previous hits and used as the new query.

example

PSI-BLAST alignment of RBP and β-lactoglobulin

- first iteration is essentially identical to a normal BLAST search
- there may be slight differences, as PSI-BLAST corrects for amino acid composition

```
Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

Query: 27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
           V+ENFD  ++ G WY + +K P      + I A +S+ E G +    K          ++
Sbjct: 33  VQENFDVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIEVLNK---------ELS 82

Query: 87  ADMVGTF---------TDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCR 137
            D  GT          ++  +PAK +++++ +          +WI+ TDY+ YA+ YSC
Sbjct: 83  PD--GTMNQVKGEAKQSNVSEPAKLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138 ----LLNLDGTCADSYSFVFSRDPNGLPPE 163
               L ++D      + ++  R+P  LPPE
Sbjct: 136 TFFWLFHVD------FFWILGRNPY-LPPE 158
```

In this example, we observe how the alignment and score for **one** hit from the entire set evolves over a number of iterations. The first E-value is $2e^{-04}$.

example

PSI-BLAST alignment of RBP and β-lactoglobulin

- second iteration: a position specific scoring matrix (profile) is computed from all significant hits of the first iteration and then used to score alignments
- in this example, the score has increased, the % id is slightly less, sequence length increased, fewer gaps in the alignment

```
Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)
```

```
Score =  140 bits (353), Expect = 1e-32
Identities = 45/176 (25%), Positives = 78/176 (43%), Gaps = 33/176 (18%)

Query: 4   VWALLLLAAWAAAERDCRVSSF--------RVKENFDKARFSGTWYAMAKKDPEGLFLQD 55
           V  L+ LA  A      +  +F          V+ENFD  ++ G WY + +K P      +
Sbjct: 2   VTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEI-EKIPASFEKGN 60

Query: 56  NIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMV---GTFTDTEDPAKFKMKYWGVASF 112
            I A +S+ E G +    K       + D   + V      ++  +PAK +++++ +
Sbjct: 61  CIQANYSLMENGNIEVLNKEL-----SPDGTMNQVKGEAKQSNVSEPAKLEVQFFPL--- 112

Query: 113 LQKGNDDHWIVDTDYDTYAVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEA 164
                +WI+ TDY+ YA+ YSC     L ++D      + ++  R+P  LPPE
Sbjct: 113 --MPPAPYWILATDYENYALVYSCTTFFWLFHVD------FFWILGRNPY-LPPET 159
```

The second E-value for the pair has decreased from $2e^{-04}$ to $2e^{-32}$. This has transformed a somewhat borderline hit to a certain homologue! If you look carefully, you will see that the detailed position of gaps has changed - consensus information can be invaluable to place gaps correctly - and the lenght of the alignment has grown considerably.

example

PSI-BLAST alignment of RBP and β-lactoglobulin

- in iteration 3, the PSSM is further refined
- score increases further, % id slightly less, fewer gaps, and lower E value (greater significance)

```
Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)
```

```
Score =  140 bits (353), Expect = 1e-32
Identities = 45/176 (25%), Positives = 78/176 (43%), Gaps = 33/176 (18%)
```

```
Score =  159 bits (404), Expect = 1e-38
Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

Query: 3   WVWALLLLAAWAAAERD--------CRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ 54
            V  L+ LA  A              +  S  V+ENFD  ++ G WY + K
Sbjct: 1   MVTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEIEKIPASFE-KG 59

Query: 55  DNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQ 114
            + I A +S+ E G +    K          V  +     ++  +PAK +++++ +
Sbjct: 60  NCIQANYSLMENGNIEVLNKELSPDGTMNQVKGE--AKQSNVSEPAKLEVQFFPL----- 112

Query: 115 KGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEA 164
                +WI+ TDY+ YA+ YSC           + ++  R+P  LPPE
Sbjct: 113 MPPAPYWILATDYENYALVYSCTTFFWL--FHVDFFWILGRNPY-LPPET 159
```

The E-value decreases further. A careful comparison of the trend of E-values can be very helpful for evaluating borderline hits. E-values of homologues almost always get dramatically smaller through the iterations. E-values of spurious hits get larger or stay approximately the same. Make it a habit to look at the **E-value trend** in questionable cases **but exclude the questionable hit from the profile** by unchecking the check-box on the search form, until you are satisfied that the sequence is a homologue after all. Getting unrelated sequences included in your profile will lead to *profile corruption*!

species selection

If you have too many hits, not all of them will be included or shown for the next iteration.

This can easily happen if e.g. if you search accross an entire kingdom (such as “Fungi”). In the era of high-throughput sequencing this is becoming quite common. But your hits are similar and redundant.

However, if you restrict your search too much, the profile may not be sensitive over larger evolutionary distances and may not find interesting *old* paralogues.

The solution is to search with a well-distributed subset of species.

You need to carefully consider the evolutionary tree to select well-spaced members for such a subset.

profile corruption

PSI-BLAST is useful to detect weak but biologically meaningful relationships between proteins.

Not all sequences can be used for a search. For example, a query with a coiled-coil motif may detect thousands of proteins that have this motif but that are not homologous.

False positives in the search can arise from the inclusion of irrelevant sequences into the profile.

Once even a single irrelevant protein sequence is included in a PSI-BLAST profile, it will not go away.

If the number of related irrelevant proteins is large, they will *take over* the profile: **profile corruption**!

profile corruption

Apply filtering of regions with low-complexity and biased composition

Adjust E-value from 0.001 (default) to a more stringent value (e.g. 0.0001).

Visually inspect the output from each iteration:

> Suspicious hits have irrelevant locations or functions, similarities only to parts of domains, fail to conserve important motifs or disulfide bridges or have poor-quality alignments with unusually high fractions of indels in the alignment.

Remove suspicious hits from inclusion by unchecking the box.

> Be conservative regarding the sequences you include, **true positives will gradually improve their E-values** with subsequent iterations, even if they are not included in the profile. You can simply include them in later iterations (or not at all, they will still be reported, even if they don't contribute to the profile). False positives will not improve significantly.

In the end, how many false positives can we expect? Unfortunately, more than we'd think. Jones & Swindells (2002)[1] have run an analysis against decoy sequences that picked up false positives in 5% of all cases, after the fifth iteration, although the E-value threshold was set to 0.001.

Even though their methodology was a bit *ad hoc* and finding false positives about 50 times more frequently than expected is not catastrophic, we must realize that protein sequences are not random strings and that significance is often hard to evaluate, because it is hard to get the *null* hypothesis right. Use caution, use common sense and in questionable cases try to use independent confirmation of homology, such as conserved binding sites or functional motifs, if possible.

[1] http://www.ncbi.nlm.nih.gov/pubmed/11893514

filtering

Filtering is used to mask parts of sequences with **low complexity** for database searches, by replacing sequence fragments with "X" (unknown residue) symbols.

Filtering is applied to the QUERY not the database

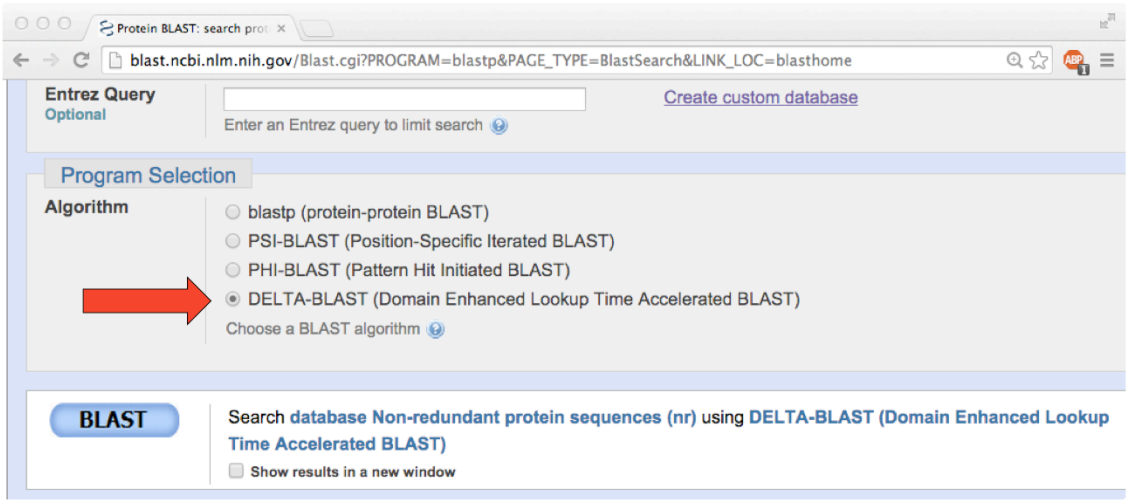
Filtering can eliminate statistically significant but biologically uninteresting reports from BLAST output (e.g. Basic, acidic, or proline rich regions, but these **may** actually be interesting, depending on your biological question.)

Personally, I turn filtering OFF by default for standard BLAST searches, only turn it ON if low-complexity regions appear to cause problems with specificity. Informed judgement is required.

However: always use filtering for PSI-BLAST (profile corruption)!

DELTA-BLAST

Invoking DELTA-BLAST is trivial:



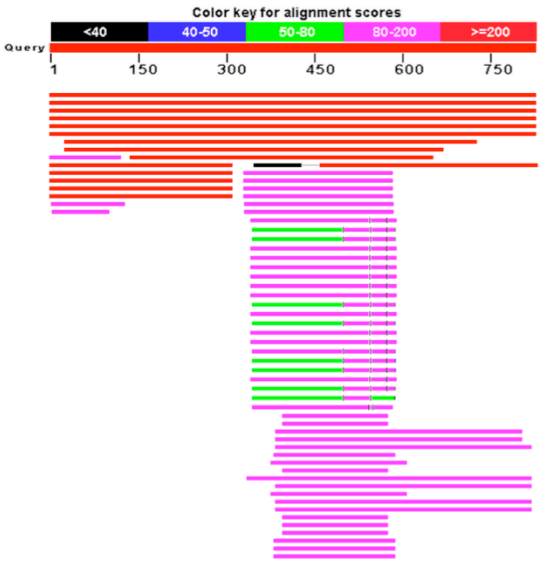
Other BLAST flavours: DELTA-BLAST:

Initiate a DELTA-BLAST search simply by choosing the option on the BLAST input form.

DELTA-BLAST

DELTA-BLAST builds a profile from the **CDD domains** in your query, and searches these in the database...

... but the results can be hard to interpret.



Mbp1 proteins in different yeast species

Proteins with irrelevant Ankyrin domains

The first interesting paralogue: Swi6

bl2seq

Access BLAST 2 Sequences by checking the box "**Align tow or more sequences**" benath the query form on the BLAST search page.

Aligns two sequences of your choice, can do different types of comparison eg. BLASTX but is NOT an optimal sequence alignment.

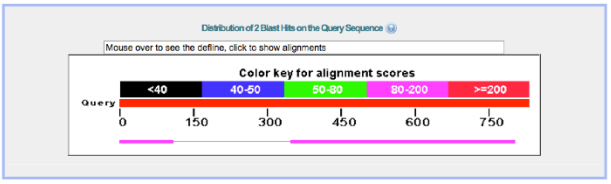
However, the **Dot Matrix View** is an excellenttool to evaluate sequence similarity between two sequences.

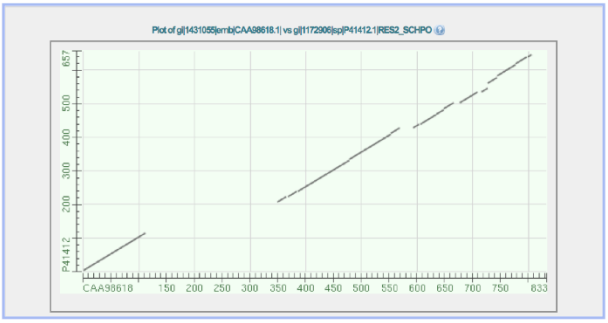
Example:

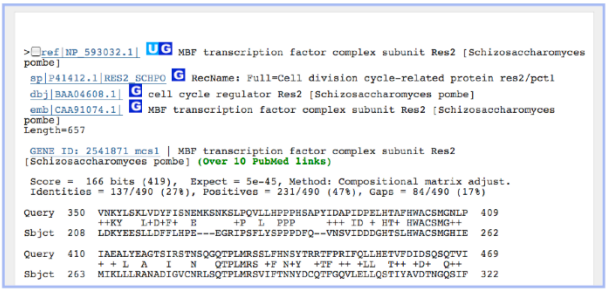
S. cerevisiae **Mbp1**

vs.

S. pombe **Res2**







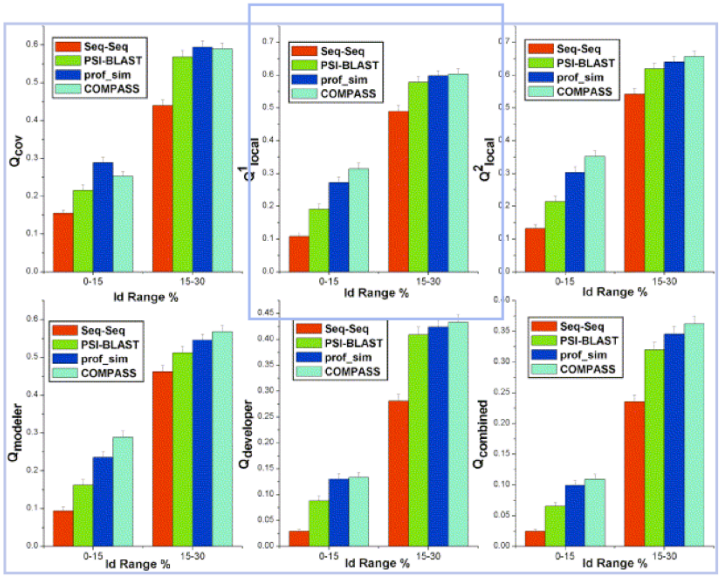
Other BLAST flavours: bl2seq

A nice extension of normal sequence alignment is the graphical view of similarities. But note that BLAST is not an **optimal** sequence alignment algorithm and I question why one would use an inferior algorithm if one has better alternatives easily available? If you plan to **use** the results, as opposed to just inspecting them visually, use EMBOSS *needle* respectively *water* instead!

That said, it is possibleto pick up good, **suboptimal** alignments – such as alignments of internal repeats, which you would not find with optimalsequence alignment methods.

profile/profile

COMPASS outperforms PSI-BLAST under all regimes tested. (Sadreyev & Grishin JMB, 2003)



http://prodata.swmed.edu/compass/compass.php

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). [...] Qcov corresponds to the portion of the length of the structural alignment that was covered by the sequence alignment, regardless of the actual accuracy. Qlocal1 and Qlocal2 correspond to the accuracy of the local prediction for only the regions that are included in the evaluated alignment. Qmodeler, Qdeveloper and Qcombined are previously suggested measures of integral accuracy of the alignment from the modeler's, developer's and combined points of view.

Is it possible to improve significantly on PSI-BLAST? Yes, Sadreyev & Grishin (2003)[1] took the idea of profile based searches further by aligning profiles of sequences against a database of profiles. The principle is the same as the "equivalence principle" for homology, sometimes we can detect distantly related homologues through a mutual similarity to an intermediate sequence.

With this and the lab's PROCAIN server[2], homology searches significantly below 20% sequence identity may be feasible.

[1] http://www.ncbi.nlm.nih.gov/pubmed/12547212

[2] http://www.ncbi.nlm.nih.gov/pubmed/19497935

See also:

http://www.ncbi.nlm.nih.gov/pubmed/19435884

http://prodata.swmed.edu/compass/compass.php

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA