Advanced Blast

The importance of similarity
Similarity derives from the same ancestral origin
If your sequences are similar,
they probably have the same ancestor,
share the same structure,
and have a similar biological function,
even when the sequences come from very different organisms.

Homologues: when two proteins or gene sequences are very similar

- 25% identity for proteins
- 75% identity for nucleotides

More information to make sure two sequences are true homologues

- the expectation value (E-value)
- the length of the similar segments
- the patterns of amino acid conservation
- the number of insertions/deletions

Blast: the most popular data-mining tool ever

Blasting protein sequences

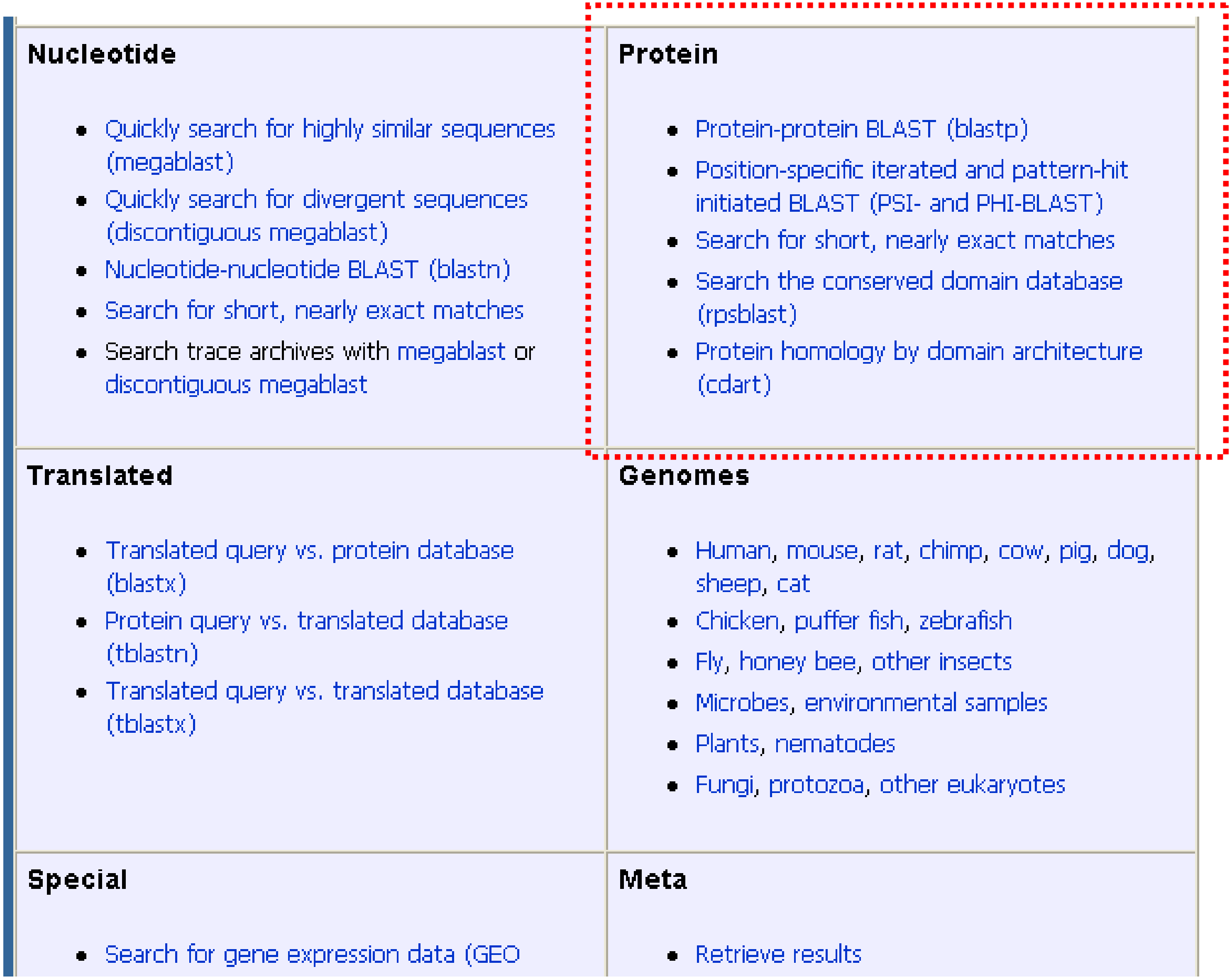
- blastp: compare a protein sequence with a protein database
  - find out something about the function of my protein
  - choose SWISS-PROT as a database

- tblastn: compare a protein sequence with a nucleotide database
  - discover new genes encoding simple proteins

Blast Net

- NCBI: www.ncbi.nlm.nih.gov
- EMBL: www.ch.embnet.org

NCBI/BLAST



Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Fly, honey bee, other insects
- Microbes, environmental samples
- Plants, nematodes
- Fungi, protozoa, other eukaryotes

Special

- Search for gene expression data (GEO

Meta

- Retrieve results

BLAST output

A graphic display: a useful tool to discover domains

A hit list:

sequence accession number
description:
bit score: the higher, the more similar. Matches below 50 bits are very unreliable

E-value: the lower, the more similar. Matches above 0.001 are very unreliable

The alignments

percent identity: either identical or similar represented with a +
length: aligned length

The parameters

default parameters

- **Peptide Sequence Databases**
  - **nr**

    All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF
  - **refseq**

    RefSeq protein sequences from NCBI's Reference Sequence Project.
  - **swissprot**

    Last major release of the SWISS-PROT protein sequence database (no updates).
  - **pat**

    Proteins from the Patent division of GenPept.
  - **pdb**

    Sequences derived from the 3-dimensional structure from Brookhaven Protein Data Bank.
  - **month**

    All new or revised GenBank CDS translation+PDB+SwissProt+PIR+PRF released in the last 30 days.
  - **env_nr**

    Protein sequences from environmental samples.

BLASTing DNA sequences

Question	Answer
Interested in noncoding DNA	blastn
Want to discover new proteins	tblastx
Discover proteins encoded in DNA	blastx
Quality of DNA (sequencing error)	blastx

Controlling BLAST

Some reasons to change BLAST default parameters

Reason	parameters to change
The sequence you are interested in Contains many identical residues; it has a biased composition	sequence filter (automatic masking)
BLAST doesn’t report any results	change the substitution matrix or the gap penalties
Your match has a borderline E-value	change the substitution matrix or the gap penalties
BLAST reports too many matches	change the database filter the reported entries by keyword increase Expect, the E-value threshold

PSI-BLAST (position specific iterated)

Finding closely related sequences

By making a new position-specific substitution matrix

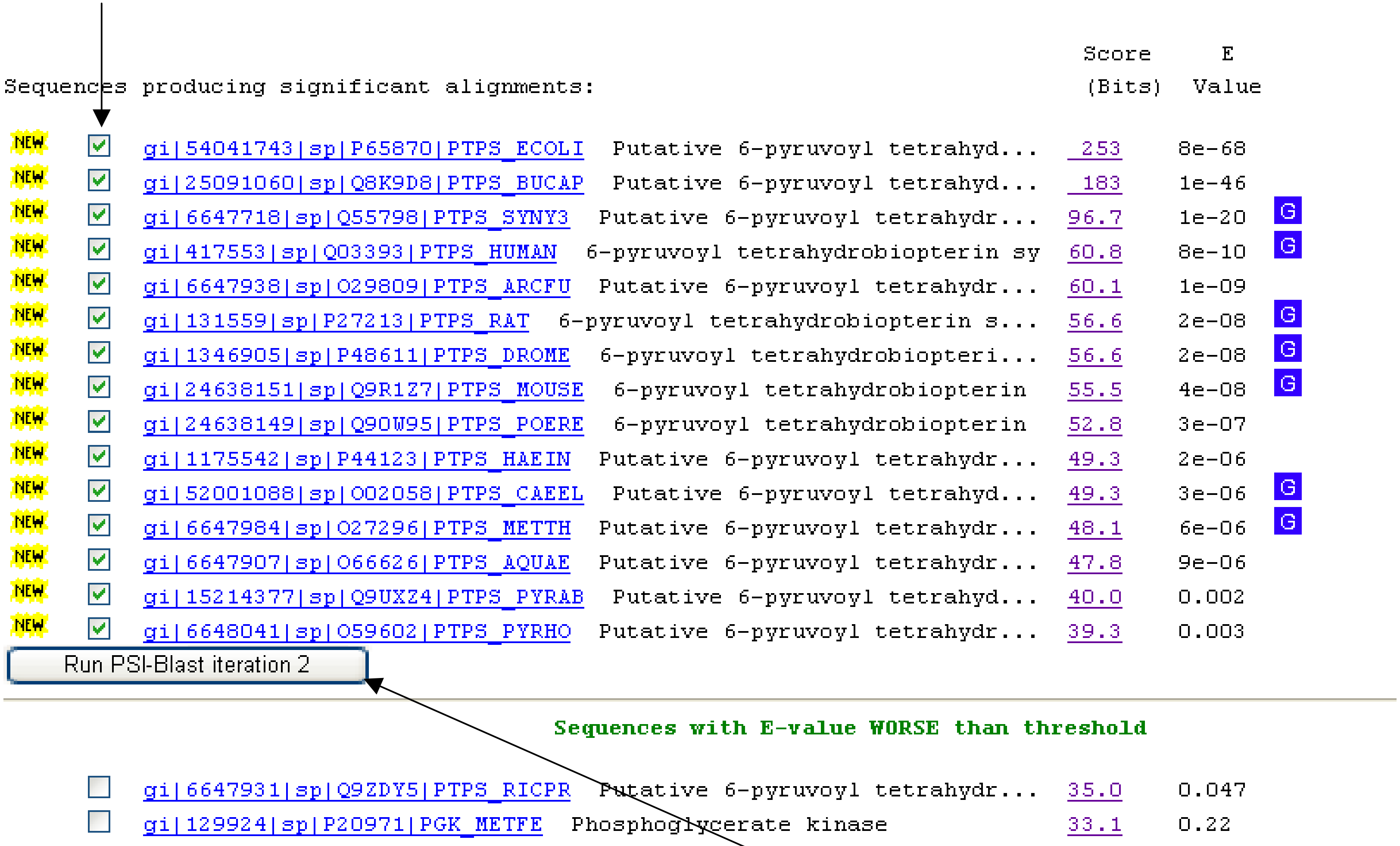
ex. Conserved Cys on position 5 and variable Cys on position 25

a new substituition matrix penalizes substitutions on position 5

while tolerates them on position 25

The main difficulty in PSI-BLAST is deciding which sequences you can keep from one iteration to the next

Check box: use the corresponding sequence to derive the position-specific matrix for the next iteration of PSI-BLAST



```
                                                                        Score    E
Sequences producing significant alignments:                             (Bits)  Value

NEW  ☑  gi|54041743|sp|P65870|PTPS_ECOLI  Putative 6-pyruvoyl tetrahyd...  253    8e-68
NEW  ☑  gi|25091060|sp|Q8K9D8|PTPS_BUCAP  Putative 6-pyruvoyl tetrahyd...  183    1e-46
NEW  ☑  gi|6647718|sp|Q55798|PTPS_SYNY3  Putative 6-pyruvoyl tetrahydr...  96.7   1e-20  G
NEW  ☑  gi|417553|sp|Q03393|PTPS_HUMAN  6-pyruvoyl tetrahydrobiopterin sy  60.8   8e-10  G
NEW  ☑  gi|6647938|sp|O29809|PTPS_ARCFU  Putative 6-pyruvoyl tetrahydr...  60.1   1e-09
NEW  ☑  gi|131559|sp|P27213|PTPS_RAT  6-pyruvoyl tetrahydrobiopterin s...  56.6   2e-08  G
NEW  ☑  gi|1346905|sp|P48611|PTPS_DROME  6-pyruvoyl tetrahydrobiopteri...  56.6   2e-08  G
NEW  ☑  gi|24638151|sp|Q9R1Z7|PTPS_MOUSE  6-pyruvoyl tetrahydrobiopterin    55.5   4e-08  G
NEW  ☑  gi|24638149|sp|Q90W95|PTPS_POERE  6-pyruvoyl tetrahydrobiopterin    52.8   3e-07
NEW  ☑  gi|1175542|sp|P44123|PTPS_HAEIN  Putative 6-pyruvoyl tetrahydr...  49.3   2e-06
NEW  ☑  gi|52001088|sp|O02058|PTPS_CAEEL  Putative 6-pyruvoyl tetrahyd...  49.3   3e-06  G
NEW  ☑  gi|6647984|sp|O27296|PTPS_METTH  Putative 6-pyruvoyl tetrahydr...  48.1   6e-06  G
NEW  ☑  gi|6647907|sp|O66626|PTPS_AQUAE  Putative 6-pyruvoyl tetrahydr...  47.8   9e-06
NEW  ☑  gi|15214377|sp|Q9UXZ4|PTPS_PYRAB  Putative 6-pyruvoyl tetrahyd...  40.0   0.002
NEW  ☑  gi|6648041|sp|O59602|PTPS_PYRHO  Putative 6-pyruvoyl tetrahydr...  39.3   0.003
```

```
                            Sequences with E-value WORSE than threshold

     □  gi|6647931|sp|Q9ZDY5|PTPS_RICPR  Putative 6-pyruvoyl tetrahydr...  35.0   0.047
     □  gi|129924|sp|P20971|PGK_METFE  Phosphoglycerate kinase              33.1   0.22
```

New: reports the hit for the first time
Green pill: used to obtain the current result

Click for the second run

Assignments

Do BLAST search with your protein
Change parameters & compare the results