

BLAST Database Searching

BME 110: CompBio Tools

Todd Lowe

April 9, 2009

Admin

- Reading:
 - Read chapter 7, and the NCBI Blast Guide and tutorial
<http://www.ncbi.nlm.nih.gov/BLAST/Why.shtml>
 - Read Chapter 8 for next class

Self-help at NCBI for BLAST

- Simple Tutorial and Guide available

<http://www.ncbi.nih.gov/Education/BLASTinfo/information3.html>

- Will use these with in-class exercises

Many helpful, detailed discussions of BLAST – read for help on homework assignments

Entire book on BLAST, available on-line within the .ucsc.edu domain:

<http://proquest.safaribooksonline.com/0596002998>

BLAST

- Basic Local Alignment Search Tool (1990)
Altschul, Gish, Miller, Myers, & Lipman

Uses short-cuts or “heuristics” to improve search speed

Like speed-reading, does not examine every nucleotide of database

However, many more choices (parameters) to make to adjust search success (over 30!!)

Varieties of BLAST

Program	Database	Query	Typical uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs, and PCR products to a genome; screening repetitive elements; cross-species sequence exploration; annotating genomic DNA; clustering sequencing reads; vector clipping
BLASTP	Protein	Protein	Identifying common regions between proteins; collecting related proteins for phylogenetic analyses
BLASTX	Protein	Nucleotide translated into protein	Finding protein-coding genes in genomic DNA; determining if a cDNA corresponds to a known protein
TBLASTN	Nucleotide translated into protein	Protein	Identifying transcripts, potentially from multiple organisms, similar to a given protein; mapping a protein to genomic DNA
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein	Cross-species gene prediction at the genome or transcript level; searching for genes missed by traditional methods or not yet in protein databases

From: *BLAST* by Joseph Bedell, Ian Korf, Mark Yandell; O'Reilly 2003

How Does it Work?

- Searches for short exact (nucleotide) or near-exact matches aka “neighborhoods” (protein) of certain “word” lengths
- Defaults:
 - Blastp: 3 amino acids
 - Blastn: 11 nucleotides
- Without an initial word match, can MISS possibly important matches
- With “seed hit”, tries to extend alignment in both directions
- A fully extended hit is an HSP (high scoring pair)

Translated BLAST

- DNA-> protein, 3 reading frames upper sequence + 3 reading frames lower sequence
- Important when you don't know or trust gene sequencing quality or annotation
- Translating in all possible frames gives additional sensitivity, avoids reading frame errors due to incorrect gene prediction or sequencing errors

Customized Applications

- Because there are so many parameters, a few web-versions of BLAST have parameters pre-set for specific applications
- MEGA BLAST – find long alignments; allows you to specify min percent identity
- “Search for short and near exact matches” – find short identical hits ~20 bases (i.e. for checking primers)
- More discussion:
<http://www.ncbi.nlm.nih.gov/BLAST/Why.shtml>

Assessing Significance

Most Basic Rules of thumb:

Two nucleotide sequences – at least 70% identical, they are likely homologous

Two protein sequences – at least 25% identical over 100 amino acid alignment

Does not take into account precise length of alignment, or number of gaps!

Not sufficient to quantitatively rank hits from a database search

Re: The “Twilight Zone”

- Less than 25% sequence identity for two protein sequences
- May still be homologous, but only similarity of 3-D protein structures can verify similar function (structural comparison tools to detect these discussed later in quarter)

Quantitative Assessment of Significance: E-values & P-values

Expressions of the same thing:

E-value : number of expected random (not biological) matches in a given db search

Examples: 0.001, 0.1, 1.0, 10, 100, 1000

P-value: probability that this hit is random (not biological)

Examples: 0.1, 0.05, 0.0001, 1×10^{-3} , 1×10^{-60}

E/P - values

Mathematical conversion between them:

$$P\text{-value} = 1 - e^{-(E\text{-value})}$$

For values < 0.01 , E-value and P-value are nearly the same

Takes into account:

- Length of sequence similarity
- Conservation of aligned nucleotides/amino acids
- Number and length of insertions and deletions
- Sizes of query sequence and database you are searching

What is Reliable?

- In biology P-value of 0.05 expect would be “good enough” (5 chances in 100 of not being correlated)
- Due to BLAST’s estimation of significance, shouldn’t blindly trust P or E values $> 1 \times 10^{-4}$
- Note: Even with a “good” E-value, the match may be between paralogs with different function! Examine alignment for local areas of high similarity (are these known domains from CDD search?)
- For good measure, I don’t have great confidence unless E value is less than 1×10^{-8}

Beware Hit Transitivity!

- “BLAST hits are not transitive, unless alignments are overlapping”

Seq1 : AAAAABBBB

Seq2 : AAAAA

Seq3 : BBBB

- Seq2 and Seq3 not necessarily homologous!

Example

- Fibrillar-like protein
 - DNA: NM_001436, Protein: NP_001427
- How “far” can we go in tree of life using nucleotide v. protein searches?
- Another query: Hox gene
 - NM_153631.2 (HOX3A)

Why would you ever use BLASTN if BlastP is more Sensitive?

- Non-translated sequences (RNA genes, promoters, etc)
- Closely related species, where you expect sequence identity > 70%

Repetitive (Low Complexity) Element Filtering

- Removes sequences that occur commonly in genomes, but do not imply functional similarity
- SINEs, LINEs, and other “selfish” DNA elements
- Simple repeats: long runs of a short (1-4) nucleotide repeats due to errors in DNA replication or structural elements (telomeres)
- Protein: polymer tracks common in trans-membrane domains, etc.
- Always use UNLESS looking for ncRNAs – can remove biologically-important hits!!

Limit Search Space

- If you only want hits to a specific species or phylogenetic group, it is ****much**** faster to only search that sub-group:
 - Under “Options for Advanced Blasting”, use “Limit by entrez query”, “or select from”:
 - “Viruses”, “Archaea”, “Bacteria”, “Mammalia”, “Homo sapiens”, etc.

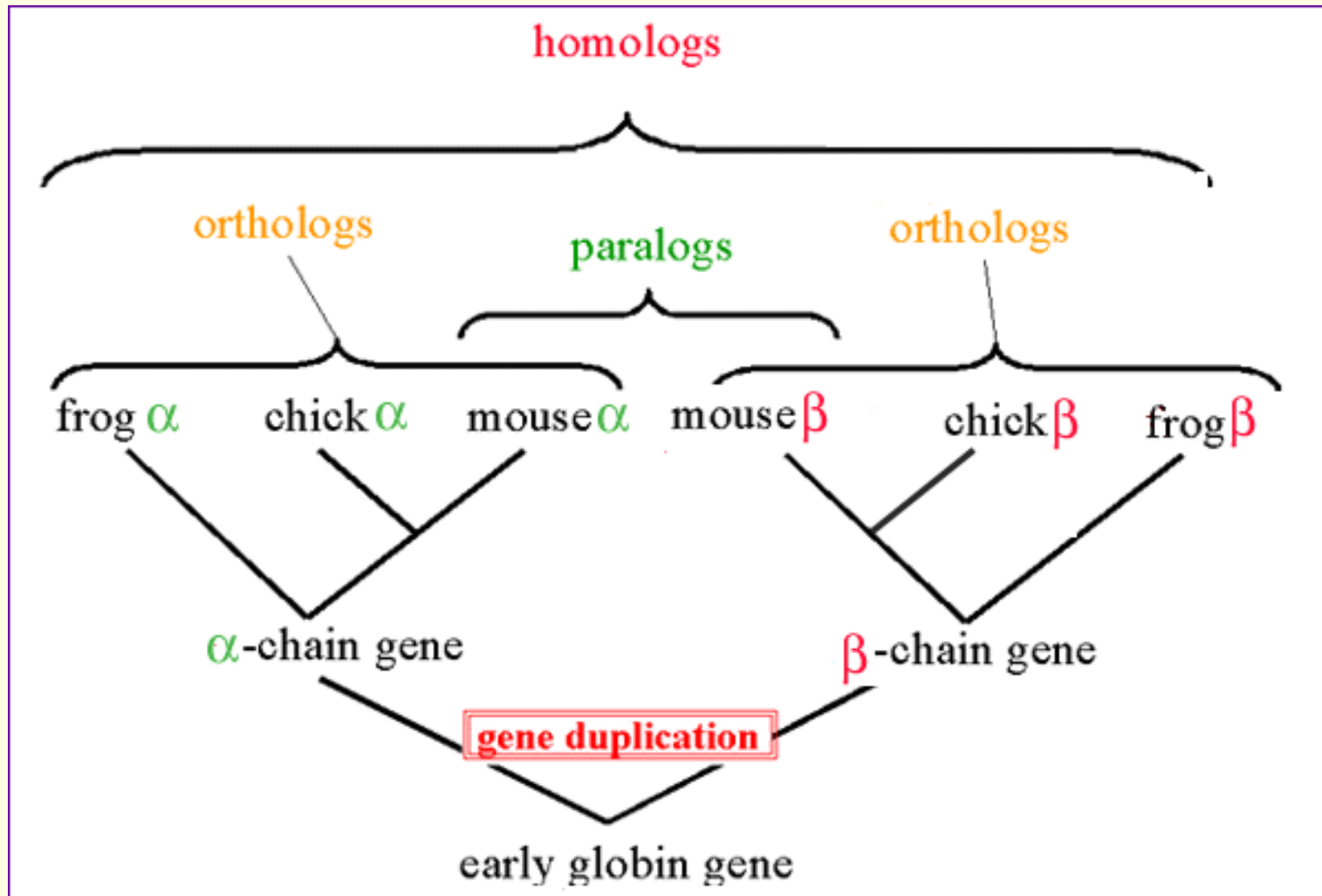
A Related Note: Homology

- Based on inference that two sequences are ancestrally derived from same molecule
- If two sequences have high *similarity*, they may be *inferred* to be homologous
- It is **WRONG** to say two sequences or genes are 80% homologous (they either are related, or they are not)

Homology: Same Function?

- Even if two sequences are ancestrally derived from same molecule, they may or may not still have the same function
 - Orthologs: homologous genes created by speciation
 - Generally implies function remains the same
 - Paralogs: homologous genes created by a gene duplication event (in same species)
 - Implies function may have changed

Homology Diagram



Source: <http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html>