

Database Searching with BLAST

Outline

- Why should we align sequences?
- Terminology
- Substitution Matrices
- BLAST programs
- Practical advice
- Interpreting results

Why align sequences?

- Sequence comparison is important for drawing functional and evolutionary inferences
- Similar sequences *may* have similar structures and functions

Evolutionary Basis

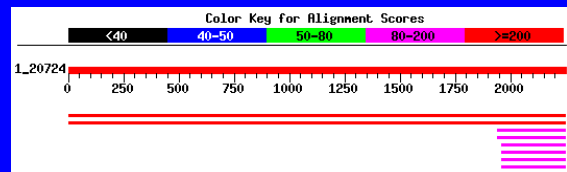
- DNA, RNA, and proteins are "molecular fossils" - they encode the history of millions of years of evolution
- During evolution, sequences accumulate random changes
- Sequences that are structurally or functionally important tend to be conserved
- Significant sequence conservation allows inference of evolutionary relationships

Homology vs. Similarity

- Homologous sequences share a common evolutionary ancestry
- Similar sequences have a high percentage of aligned residues with similar properties
- We can *infer* homology from similarity, but we can't prove it

Global vs. Local Alignment

- Global alignment - find the best alignment between entire sequences
- Local alignment - find the best alignment between *parts* of sequences



Why local alignments?

- Ignore stretches of non-coding DNA
 - More likely to contain mutations
 - Local alignment between two coding regions of DNA is likely to be between two exons
- To locate protein domains or motifs

Gaps

- Gaps in a sequence alignment represent in/dels in the evolutionary history
- A single in/del event may result in a gap of one character or many characters
- When scoring alignments you can choose separate gap existence and gap extension penalties
 - Bigger gaps don't necessarily mean sequences are farther apart evolutionarily

Substitution Matrices

- Not all mutations are equal in proteins
- Some amino acids are more exchangeable than others (Ser and Thr are similar, Ala and Trp are not)
- A substitution matrix defines different scores for each possible substitution
- Generally not applied to DNA sequences

BLOSUM and PAM

- For our purposes, we just need to know the basics about BLOSUM and PAM matrices
- Higher BLOSUM numbers mean we want more similar sequences
 - BLOSUM80 is better than BLOSUM45 when searching for highly similar sequences
- Higher PAM numbers mean we want less similar sequences

Database Searching

- Given a sequence (either DNA or protein), find similar sequences in a database
- Simple approach – compare the query sequence with all sequences in the database
- PROBLEM – sequence databases are HUGE and this takes a long time

BLAST

- Basic Local Alignment Search Tool
- The most used bioinformatics program - NCBI servers handle more than 100,000 BLAST searches per day!
- Allows fast searching of large databases by NOT comparing the query sequence with all sequences in the database
- BLAST is NOT guaranteed to find the best match

Different BLAST "Flavors"

- **BLASTN** - DNA seq against a DNA DB
- **BLASTP** - protein seq against protein DB
- **BLASTX** - 6 frame translated DNA seq against a protein DB
- **TBLASTN** - protein seq against 6 frame translated DNA DB
- **TBLASTX** - 6 frame translated DNA seq against 6 frame translated DNA DB

Different BLAST "Flavors"

- **PSI-BLAST** - protein profile against protein DB
- **PHI-BLAST** - protein pattern against protein DB
- **MEGA-BLAST** - optimized for highly similar sequences

Which BLAST program should I use?

- Start simple - if you have a DNA sequence, use BLASTN, if you have a protein sequence, use BLASTP
- Chances are pretty good that your DNA or protein sequence is already in the database (or at least a very similar sequence is there)

What if my search returns nothing?

- Try changing the algorithm parameters
- Use a different substitution matrix (lower BLOSUM number, higher PAM number)
- Change the gap costs - lower gap costs may find more hits
- Change the word size - smaller word size may find more hits

Still nothing?

- Use one of the translated BLAST programs
- Use PSI-BLAST - may be able to find more distantly related sequences

How can I tell if my BLAST results are significant?

- E-value - the lower the better
- E-value definition - the number of alignments with this score or better expected to occur in this size database by chance
- Example - an E-value of 2 means that you can expect 2 alignments this good or better to occur *just by chance*, NOT because the sequences are related

But my protein doesn't look very similar to the BLAST hits

- Greater than 25-30% identity means they are likely related
- 15-25% identity is known as the "twilight zone" - may or may not be related
- Less than 15% identity means they are NOT likely to be related

BLAST Output

- [Sample BLAST results](#)

- <http://dobbslab.gdcb.iastate.edu/BCB590>