Orthologous Gene Finding

August 7,2007

Joe Rodriguez
BaRC

Purpose

- Find gene in another genome
- Find identifying features of a group of orthologs or family members
- Determine evolutionary relationships between orthologs or gene families

Viewpoints

- Two ways to view evolution of a pair of orthologs
  - Both genes arose from common ancestor
    - Similar gene structure, secondary structure, common sequence characteristics
  - Genes evolved in parallel – Functional Orthologs
    - Sequence identity extremely low,
      - however secondary structure similar
      - conserved motifs

Twilight Zone…

1: Twilight zone of protein sequence alignments. Rost B. Protein Eng. 1999 Feb;12(2):85-94. From the Abstract:

- Sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity is high (>40% for long alignments). The signal gets blurred in the twilight zone of 20-35% sequence identity.

Low % ID != not ortholog

- If we assume that two organisms share ancestor
  - Mutations occurring at random over time means segments of high identity and low identity

Q: What if time since organisms split = VERY long time?

A: We need a more sensitive way of finding a link between two sequences than just sequence identity.

Ortholog Search Overview

- Using Literature Search –
  - Using resources such as pubmed, find out conserved residues, motifs, domains, functional residues
- Who am I? Introspective Analysis –
  - Find out information about the sequence itself as well as traits it shares with family members
- Database Search –
  - Search Protein databases using Pfam, and Blast to identify ortholog candidates, even if partial alignment, %ID, or unassembled genome
- Comparison
  - Create an multiple sequence alignment to group your sequences together
  - Use gene characteristics from first two steps above to biologically align your sequences.

Who am I: Gene Characteristics

- What do we know about the gene?
  - High similarity between:
    - Other Species
    - Family members
  - conserved residues, motifs
  - Protein domains

Protein Domains

- Proteins can have different combinations of functional domains

- We can characterize protein function based on domains, and therefore use these features to search for orthologs, or family members

  How? Pfam.

Pfam

- Database of close to 9000 protein domains from several organisms.
- Updated every 6 months or so
- Use a statistical approach to determine the likelihood that a sequence contains a protein domain.
  - With Hmmer we can scan sequences using a Pfam profile

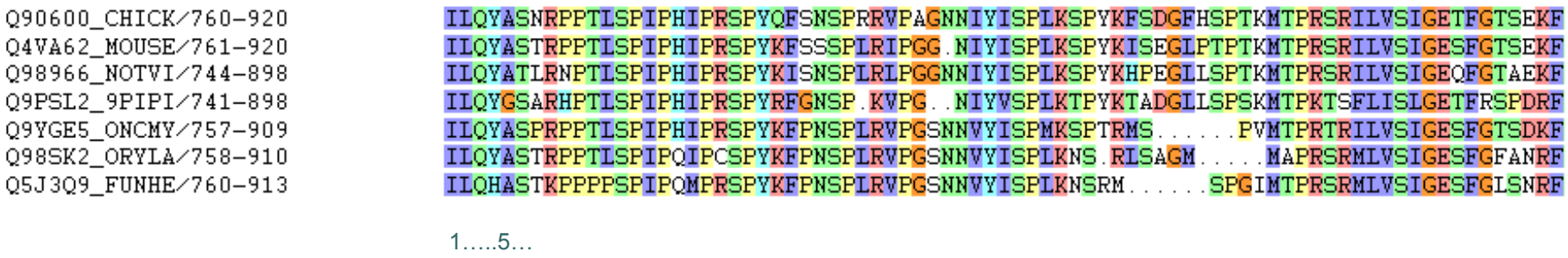
What's a Domain Profile?

- In essence, the curators at Pfam have:
  - Retrieved all known sequences identified as a particular domain (seed sequences)
  - Created a multiple sequence alignment
  - Calculated the frequency of an amino acid at a position in the domain.

This allows you to take an identified domain, and add it to a profile without considering sequence identity

Example seed sequence MSA

Rb C-terminal domain



Hmmer Usage

- Web Tool: http://www.sanger.ac.uk/Software/Pfam/search.shtml
  - Demo
- Command Line
  - Search One Profile
    1. Find exact Pfam domain name Case sensitive
    2. Hmmfetch Pfam_ls “Shugoshin_C” > mydatabase.txt
    3. Hmmsearch –cut_tc mydatabase.txt sequencefile > output.txt
  - Search More than One Profile
    1. Hmmpfam –cut_tc Pfam_ls sequencefile > output.txt

Ortholog Search Workflow

First Step:

- Literature search – Motifs/Residues, and sequence retrieval.
- Run Pfam/BlastP against family sequences
- Create MSA and add protein domain info and literature info

Ortholog Search Workflow cont

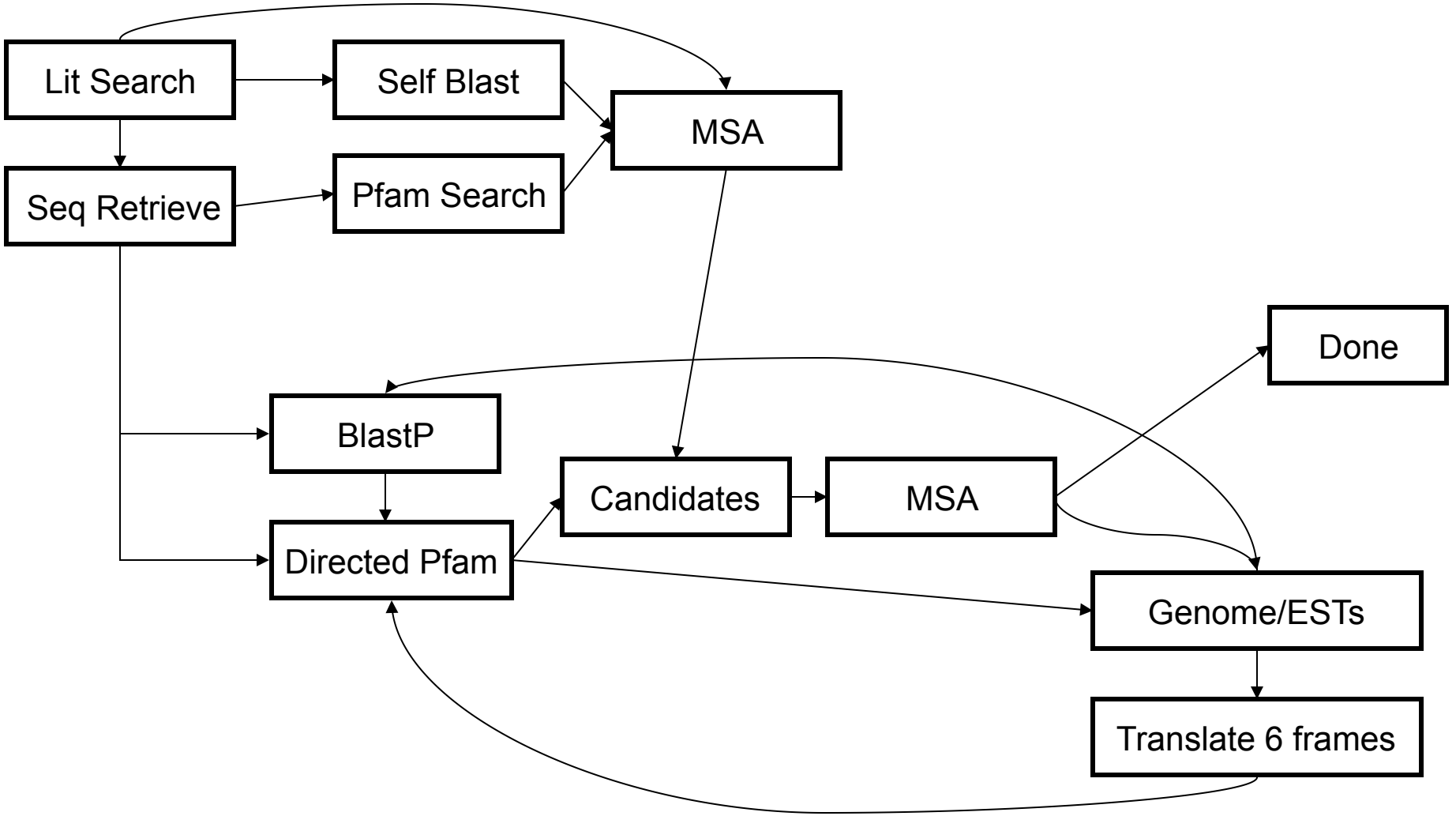
Second Step

- Blastp gene against
  - Proteome of Interest
    - Candidates -> Create MSA
- Run a directed Pfam against proteome
  - See if candidates share similar domains to gene of interest
  - Caveat: Make sure protein domain is NOT entire gene. If multiple domains make up the protein domain, flag this. Try using nondomain sequence to give more weight.

- If no candidates/or want to be thorough
  - Run Pfam against all 6 frames of translated genome/est DB
  - Run Blastp against EST/genome DB

- Once we have candidates:
  - Create a multiple sequence alignment via Clustalw.
  - Open the .aln file with Jalview and use protein domain/lit info

Ortholog Search Workflow



Jalview

- Why Jalview?
  - Jalview allows you to edit the multiple sequence alignment.
    - Why??? Computational algorithms are too fitted for high sequence similarity/scoring. Even at high sequence similarity we need to make sure the biology makes sense, even if scoring scheme suggests otherwise.
- Using what we found from pfam and literature search, we should attempt to see if we can find these features in our MSA. If not we try manually editing the software.
- The coloring of the editor is helpful

Jalview

- Inputs:
  - Fasta
  - clipboard
  - Clustal .aln files
- Output
  - Image files (not so nice for publication) use clustalx to export to postscript
- Example demo
- http://www.jalview.org

In a pickle

- Family of proteins with same protein domains and high sequence similarity makes it difficult to distinguish
- Sequence comprising of all domains makes it difficult to distinguish
- Very little sequence similarity or candidates
  - What next? Search the Genome using blastp and Pfam.
- Two protein candidates in desired organism, sharing sequence similarity to different parts of gene of interest, however very little sequence similarity between the candidates

Conclusion

- Ortholog searching can be a time consuming and iterative process (and subjective)
- Multiple sequence alignment editors are very useful in determining the best biologically meaningful alignment
- You may want to try looking at secondary gene structure, however these algorithms rely on multiple sequences of the same gene/orthologs.

What's next? The study of bends/conformation motifs in RNA to help align sequences with low percent ID.