SYNERGY2 *C. glabrata* Orthology Discrepancy Analysis

Author: Hubert Kicinski

Affiliation: Dr. Bin Z. He Lab @ The University of Iowa

Date: April 19, 2025

Contact: hkicinski@uiowa.edu

Overview

This document presents the findings from a systematic investigation into the discrepancy in orthology assignments for *Candida glabrata* between SYNERGY2 and two reference databases (YGOB/CGOB). Our previous orthology comparison analysis across four yeast species revealed a consistent pattern where *C. glabrata* showed high precision (0.71-0.87) but low recall (0.33-0.34) across all comparisons. This pattern suggests that while SYNERGY2 makes accurate ortholog predictions for *C. glabrata*, it misses many orthologs that are present in the reference databases.

Observed Discrepancy

In our comprehensive orthology detection comparison, we observed that:

Strong performance (F1 > 0.85) for three species pairs:

- *C. albicans - S. cerevisiae*
- *K. lactis - S. cerevisiae*
- *C. albicans - K. lactis*

Moderate performance (F1 ~ 0.45-0.49) for all *C. glabrata* comparisons:

- *C. glabrata - K. lactis*
- *C. albicans - C. glabrata*
- *C. glabrata - S. cerevisiae*

C. glabrata pattern: Consistently showing high precision (0.71-0.87) but low recall (0.33-0.34)

Research Question

Why does C. glabrata show significantly lower recall in SYNERGY2's orthology predictions compared to YGOB/CGOB reference data, while maintaining high precision and demonstrating different behavior than other yeast species?

Methods

Data Sources

SYNERGY2 Output Files:

- `clust_to_trans.txt`: Mapping between cluster IDs and gene IDs

- `root.pep`: Protein sequences used in SYNERGY2 analysis
- `final_clusters.txt`: Final ortholog cluster assignments

Reference Data:

- YGOB/CGOB ortholog pair files for all species pairs
- Focus on *C. glabrata* - *S. cerevisiae* pairs for in-depth analysis

Analysis Approach

1. **Identification of Concordant and Discordant Pairs**:
   - **Concordant pairs**: Gene pairs assigned as orthologs in both SYNERGY2 and YGOB/CGOB
   - **Discordant pairs**: Gene pairs assigned as orthologs in YGOB/CGOB but placed in different clusters by SYNERGY2
2. **Extraction of Protein Sequences**:
   - Retrieved protein sequences from SYNERGY2's `root.pep` file
   - Mapped gene IDs to sequence positions using `clust_to_trans.txt`
3. **Sequence Similarity Analysis**:
   - Performed BLASTp analysis with sensitive parameters for both concordant and discordant pairs
   - Parameters included low word size (2), high E-value threshold (10), and disabled filtering
   - Analyzed percent identity, query coverage, E-value, and bit score
4. **Comparative Analysis**:
   - Compared sequence similarity metrics between concordant and discordant sets
   - Identified trends and potential thresholds used by SYNERGY2 for orthology assignment

Results

Concordant Pairs Analysis

- 18 pairs analyzed
- 7 pairs (38.9%) showed statistically significant similarity (E-value < 0.01)
- Average percent identity: 49.92%
- Average query coverage: 38.56%
- Highest identity pair: CAGL0I08591g vs YDR085C (97.51% identity, 100% coverage)
- Lowest identity pair: CAGL0L06314g vs YDR101C (27.10% identity)

Notable examples:

- CAGL0G05071g vs YDR062W: 81.56% identity, 100% coverage
- CAGL0B00924g vs YCL029C: 78.38% identity, 65% coverage
- CAGL0G08063g vs YDR092W: 34.48% identity, 94% coverage

Discordant Pairs Analysis

- 15 pairs analyzed in the expanded analysis

- Only 1 pair (6.7%) showed statistically significant similarity (E-value < 0.01)
- Average percent identity: 35.26%
- Average query coverage: 29.93%
- Highest identity pair: CAGL0E00935g vs YDR167W (66.67% identity but only 8% coverage)
- Lowest identity pair: CAGL0F05753g vs YDR182W (16.67% identity)

Notable examples:

- CAGL0I03608g vs YDL153C: E-value of 0.004 (only statistically significant pair)
- CAGL0F05863g vs YDR178W: 25% identity, 97% coverage, E-value of 0.039
- CAGL0J06374g vs YDL130W: 34.04% identity, 58% coverage, E-value of 0.028

Key Differences Between Concordant and Discordant Pairs

1. (**Statistical Significance**: 38.9% of concordant pairs showed statistically significant similarity (E-value < 0.01) compared to only 6.7% of discordant pairs.
2. **Sequence Coverage**: Concordant pairs had higher average query coverage (38.56% vs. 29.93%), indicating more extensive regions of similarity.
3. **Sequence Identity**: Concordant pairs showed higher average percent identity (49.92% vs. 35.26%).
4. **E-value Distribution**: The best E-value in discordant pairs was 0.004, suggesting a potential threshold around 0.01 for SYNERGY2's sequence similarity criterion.

Discussion

SYNERGY2's Implicit Sequence Similarity Threshold

The results strongly suggest that SYNERGY2 employs an implicit threshold for sequence similarity when determining orthology. The data points to an approximate E-value threshold of 0.01-0.05, below which genes are considered candidates for the same orthology cluster. This explains why many gene pairs considered orthologs by YGOB/CGOB, but with limited sequence similarity, are placed in different clusters by SYNERGY2.

Non-sequence Factors in Orthology Assignment

The concordant pairs analysis reveals that SYNERGY2 can place genes in the same cluster despite having relatively low sequence similarity in some cases. This suggests that synteny (genomic context) plays a supporting role in SYNERGY2's assignments, especially for borderline cases.

Biological Factors Affecting *C. glabrata* Orthology Detection

Several evolutionary characteristics unique to *C. glabrata* may explain the observed pattern:

Accelerated Sequence Evolution: *C. glabrata* has experienced more rapid sequence evolution compared to other yeast species. This results in greater divergence from orthologous genes, pushing many below SYNERGY2's detection threshold. Genomic Restructuring: Following the whole genome duplication event, *C. glabrata* underwent significant genomic restructuring, including extensive gene loss and chromosomal rearrangements. Reductive Evolution: As a human pathogen, *C. glabrata* has undergone reductive evolution, losing genes that were no longer necessary for its niche. This has resulted in a more streamlined genome with greater divergence from ancestral yeast species. Different Whole Genome Duplication Resolution: *C. glabrata* and *S. cerevisiae* resolved the whole genome duplication differently, retaining different paralogs in many cases, which makes one-to-one orthology assignments particularly challenging.

Methodological Differences in Orthology Assignment

The discrepancy also highlights fundamental differences in orthology assignment approaches:

1. **SYNERGY2 Approach**:
    - Places strong emphasis on detectable sequence similarity
    - Requires statistical significance in sequence alignment (approximately E-value < 0.05)
    - Uses synteny as supporting evidence but not as a primary criterion in the absence of sequence similarity
    - Prioritizes specificity (high confidence) over sensitivity (detecting all possible orthologs)
2. **YGOB/CGOB Approach**:
    - Incorporates additional evidence beyond sequence similarity
    - Relies more heavily on synteny and gene order
    - Includes manual curation based on broader evolutionary evidence
    - Accepts pairs with minimal or no detectable sequence similarity as orthologs

Conclusions

The discrepancy in C. glabrata orthology assignments between SYNERGY2 and YGOB/CGOB seems to stemmm from a combination of biological factors and methodological differences, not technical errors in analysis (as mentionded). SYNERGY2 requires a minimum threshold of sequence similarity (approximately E-value < 0.05) to place genes in the same orthology cluster. Many *C. glabrata* genes fail to meet this threshold due to rapid sequence evolution. The pattern of high precision but low recall for *C. glabrata* reflects SYNERGY2's conservative approach to orthology assignment, which prioritizes confidence (specificity) over completeness (sensitivity). YGOB/CGOB uses additional criteria beyond sequence similarity, including synteny and manual curation, to identify orthologs when sequence similarity is low. Neither approach is inherently "wrong" - they represent different trade-offs between specificity and sensitivity in ortholog detection. This analysis provides valuable insights into the unique evolutionary characteristics of *C. glabrata* and highlights the challenges in orthology detection.

Future Directions

- Synteny Analysis: Investigate whether discordant pairs maintain syntenic relationships despite low sequence similarity. Functional Analysis: Determine if specific functional categories of genes show greater discordance.
- Parameter Testing: Experiment with SYNERGY2 parameters to adjust the relative weights of sequence similarity vs. synteny.

(and hopefully more)