PathBLAST: a tool for alignment of protein interaction networks

Ajay G.H

M.S Computational Biology

Introduction

- Protein network alignment tool
- Comparison of protein networks between two different organisms
- Goal:
  - To identify:
    - Similar protein pathways
    - Conserved protein complexes

What are the challenges in PPI?

- High dimensionality of the data
- Complexity of information
- How to separate true PPI or protein DNA interactions from false positives?
- How to functionally annotate the networks?
- How to effectively use the large network information in building or understanding different models?

Why Network Alignment?

- Excellent way of cross species comparison
- Comparison of two networks
  - Identification of conserved signal
  - High evolutionary significance
  - Identification of variations in the pathways

Description

- PathBLAST:
  - Query pathway aligned against target network
  - Goal:
    - Similar pathways in target network
  - Query:
    - 2-5 proteins long
    - Protein sequences or ID’s
  - Target:
    - Complex protein networks
    - Currently 7 different targets
    - DIP database

Targets

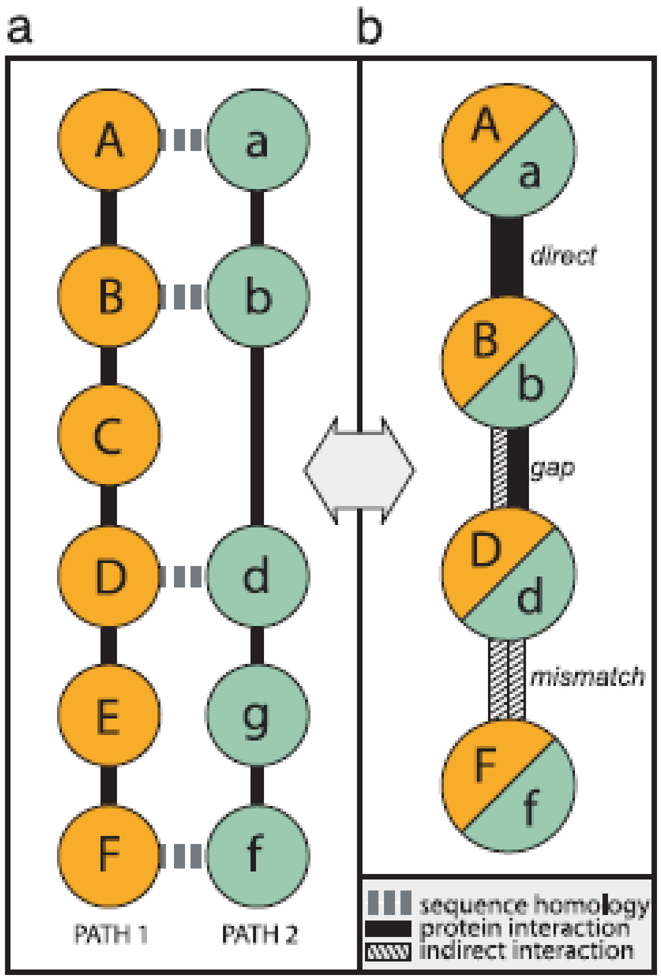
- *Saccharomyces cerevisiae*
- *Helicobacter pylori*
- *Escherichia coli*
- *Caenorhabditis elegans*
- *Drosophila melanogaster*
- *Mus musculus*
- *Homo sapiens*
- Excellent cross species information
  - Coli to humans

Alignment

- PathBLAST searches for high scoring pathway alignment
- Proteins of the first path are paired with putative orthologs occurring in the same order in the second path
- Incorporation of evolutionary significance?
  - Gaps and mismatches are allowed

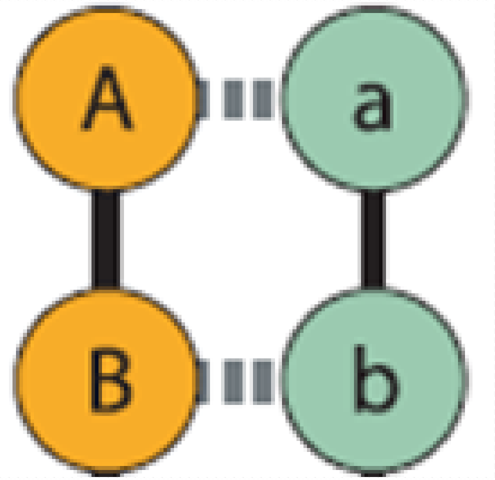
Gaps and Mismatches

- Gap:
  - Common protein
  - Interactions in one path may skip over a protein in other pathway introducing a gap
- Mismatch:
  - Proteins in similar position that are dissimilar in sequence introduce mismatches



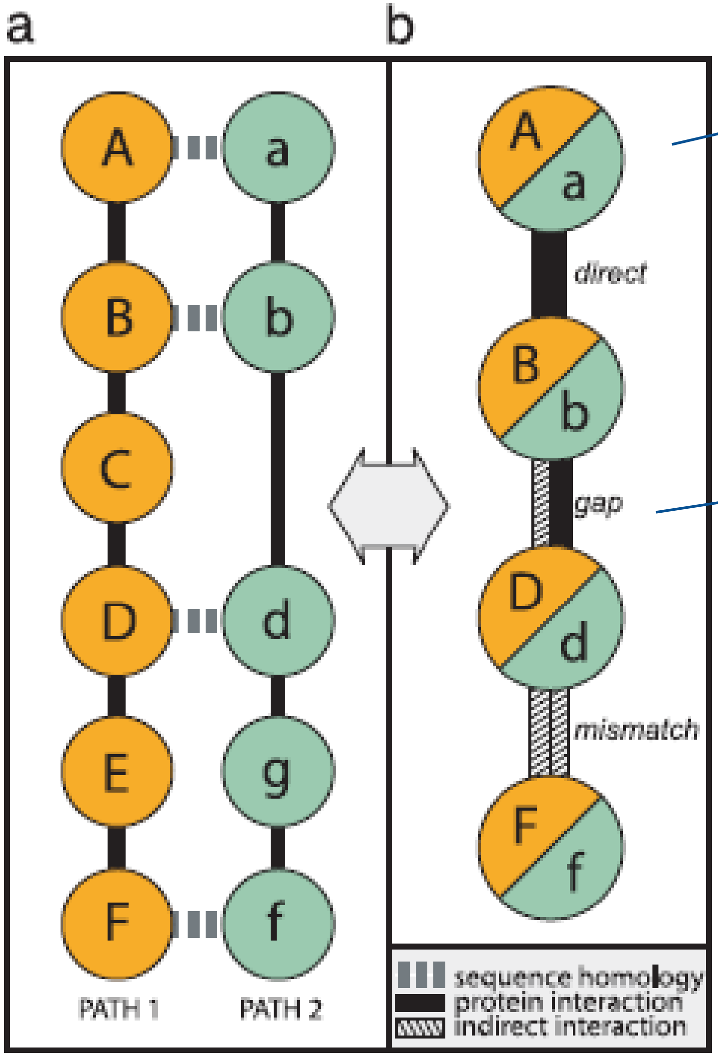
BLAST in PathBLAST

- Identification of homologous proteins between querry and target
- Based on e-value
- Narrows down the unwanted interactions that usually have a high e-value



Algorithm

- Pathways are combined as a global alignment graph
- Each node represents a homologous protein pair
- Links represent protein interaction relationships
- Three types of relationships:
  - Direct interaction
  - Gap (one interaction is indirect)
  - mismatch (both interactions are indirect)



Homologous protein pairs

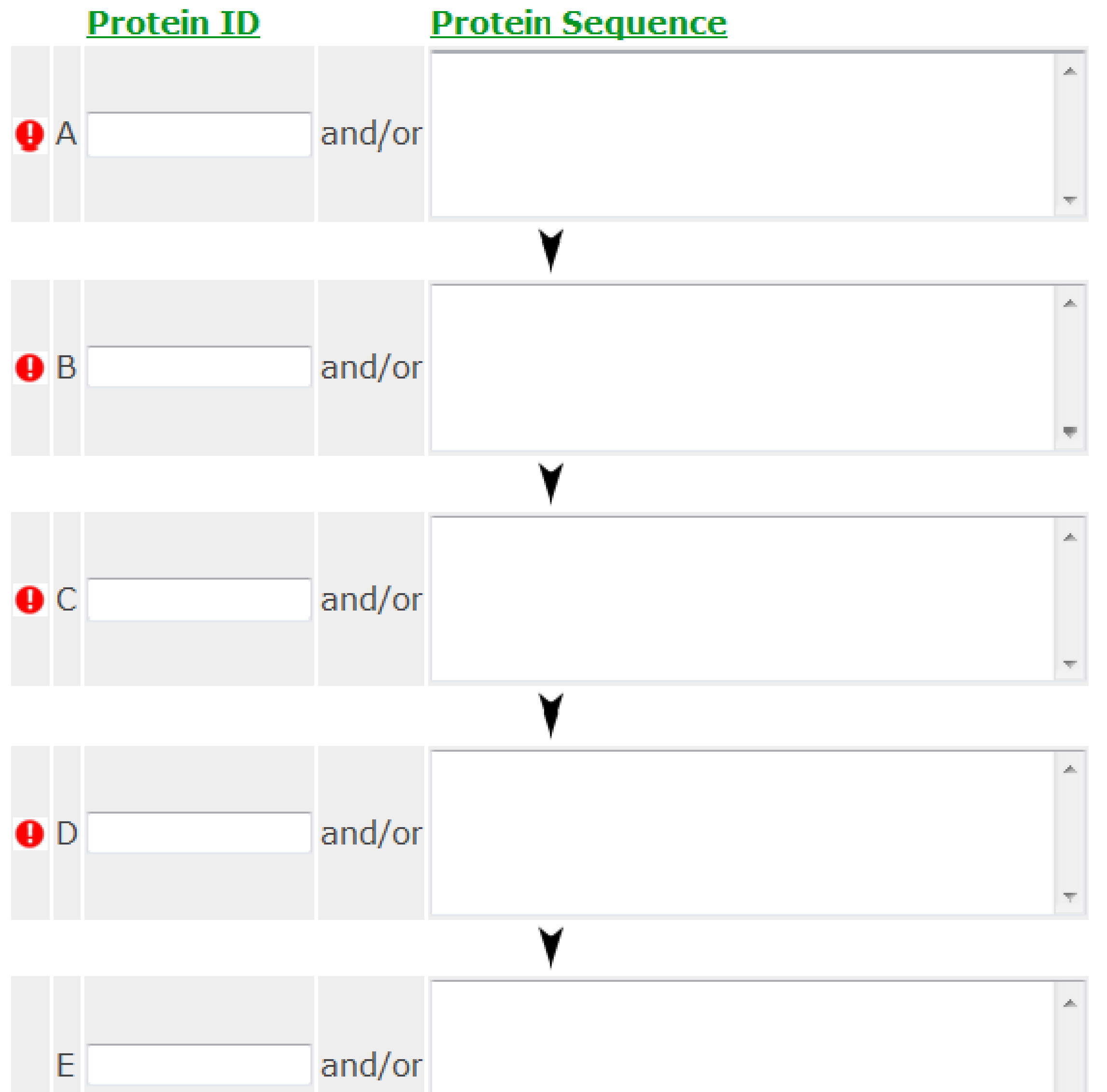
Interactions

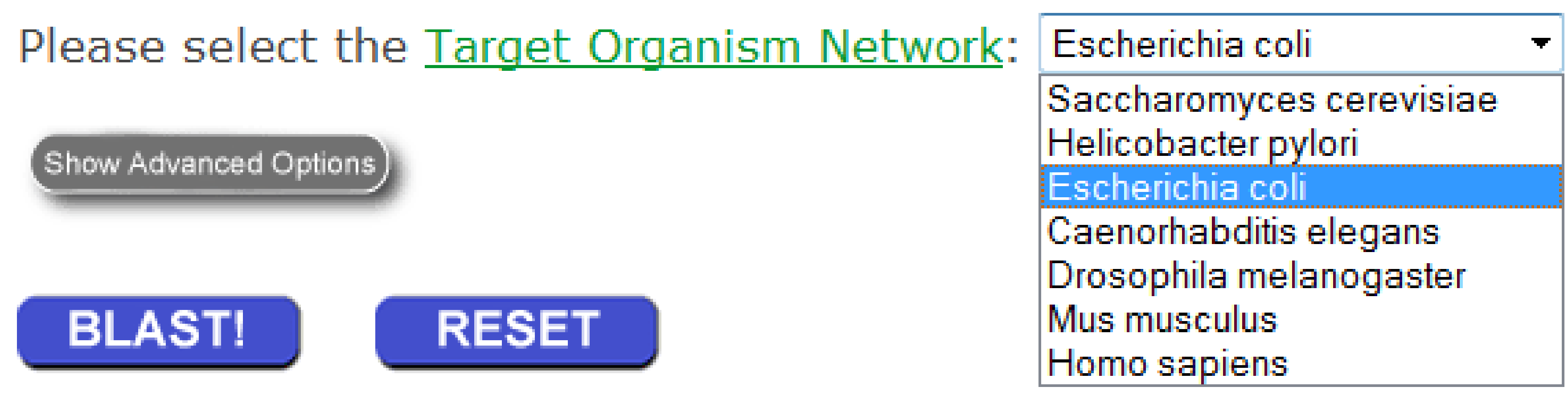
score

$$S(P) = \sum_{v \in P} \log_{10} \frac{p(v)}{p_{\text{random}}} + \sum_{e \in P} \log_{10} \frac{q(e)}{q_{\text{random}}},$$

- where p(v) is the probability of true homology within the protein pair represented by v
- q(e) is the probability that the protein–protein interactions represented by e are real, i.e., not false-positive errors.
- The background probabilities prandom and qrandom are the expected values of p(v) and q(e) over all vertices and edges in the global alignment graph.

Web interface

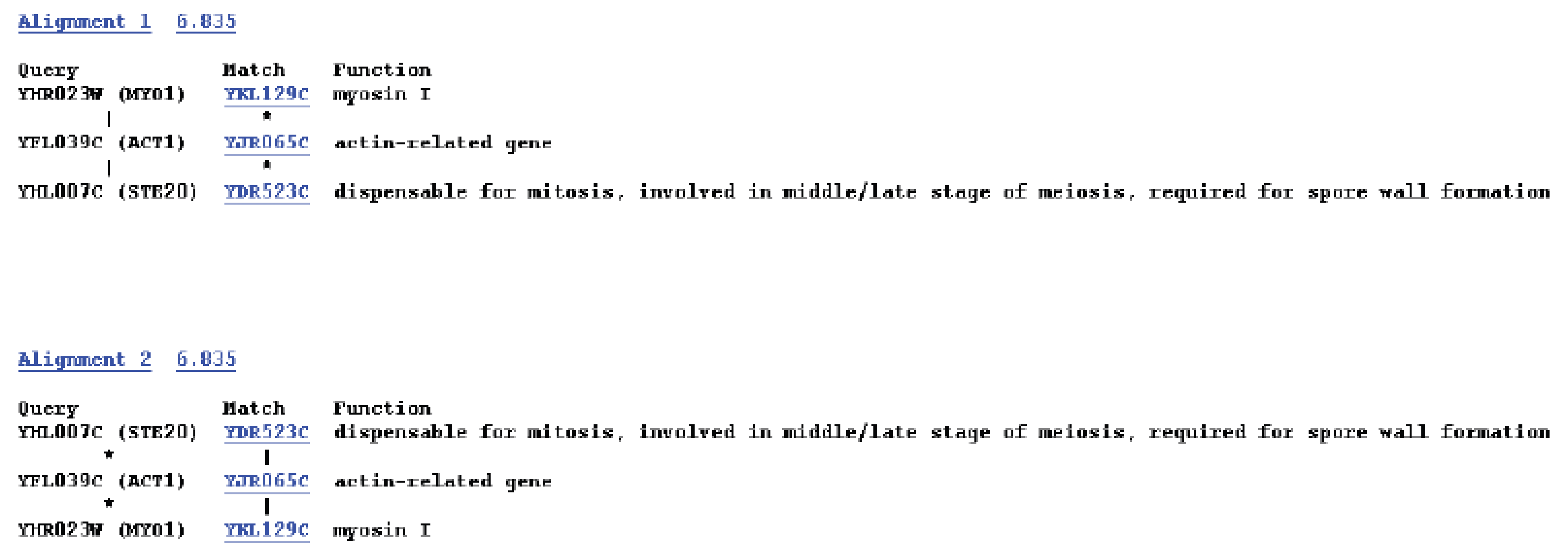




Please select the Target Organism Network: Escherichia coli

Results

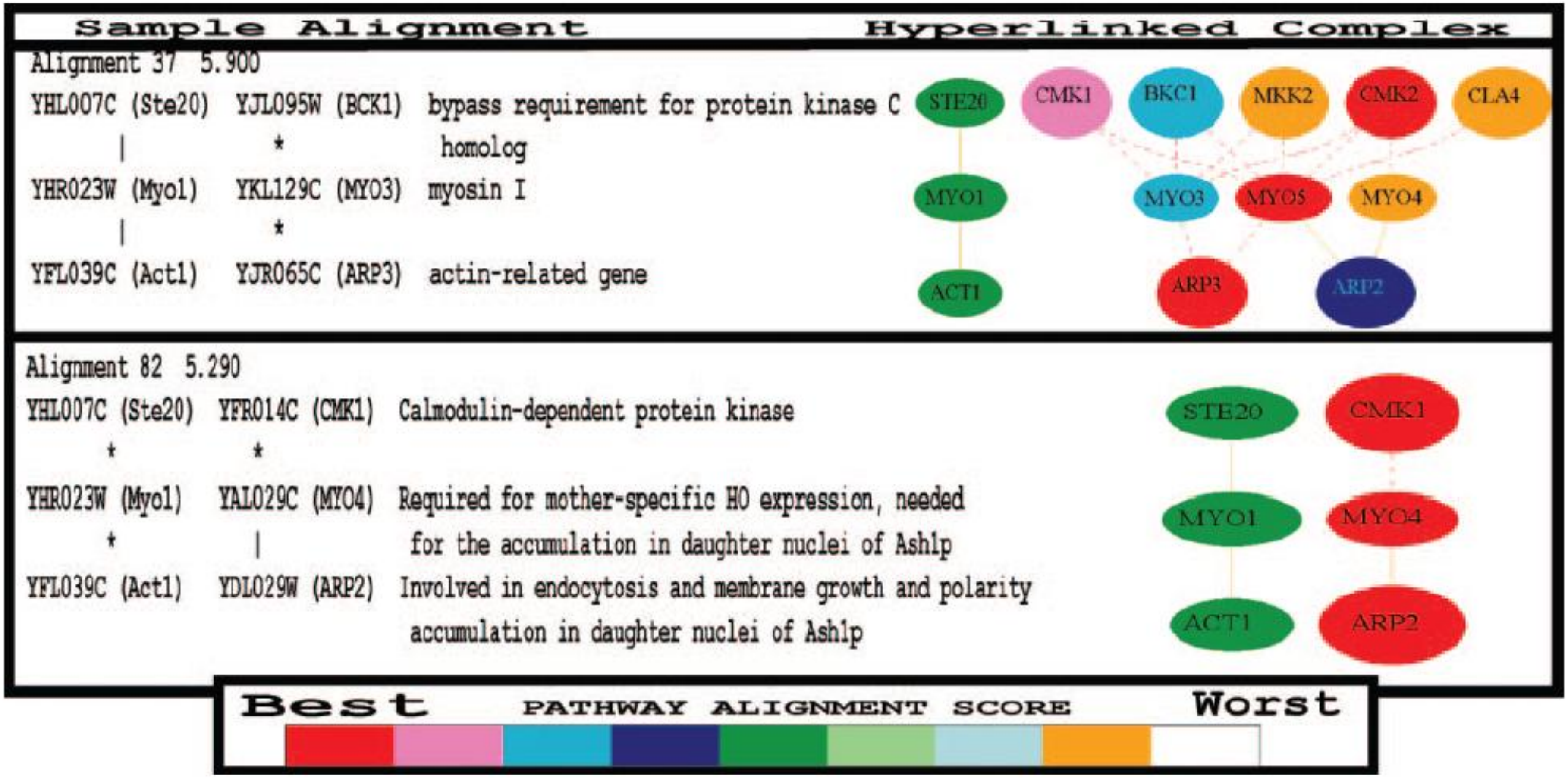
- Text and graph : useful hyperlinking
- Text results



```
Alignment 1  6.835

Query            Match    Function
YHR023W (MYO1)   YKL129c  myosin I
       |             *
YFL039c (ACT1)   YJR065c  actin-related gene
       |             *
YHL007c (STE20)  YDR523c  dispensable for mitosis, involved in middle/late stage of meiosis, required for spore wall formation



Alignment 2  6.835

Query            Match    Function
YHL007c (STE20)  YDR523c  dispensable for mitosis, involved in middle/late stage of meiosis, required for spore wall formation
       *             |
YFL039c (ACT1)   YJR065c  actin-related gene
       *             |
YHR023W (MYO1)   YKL129c  myosin I
```

- Graph results

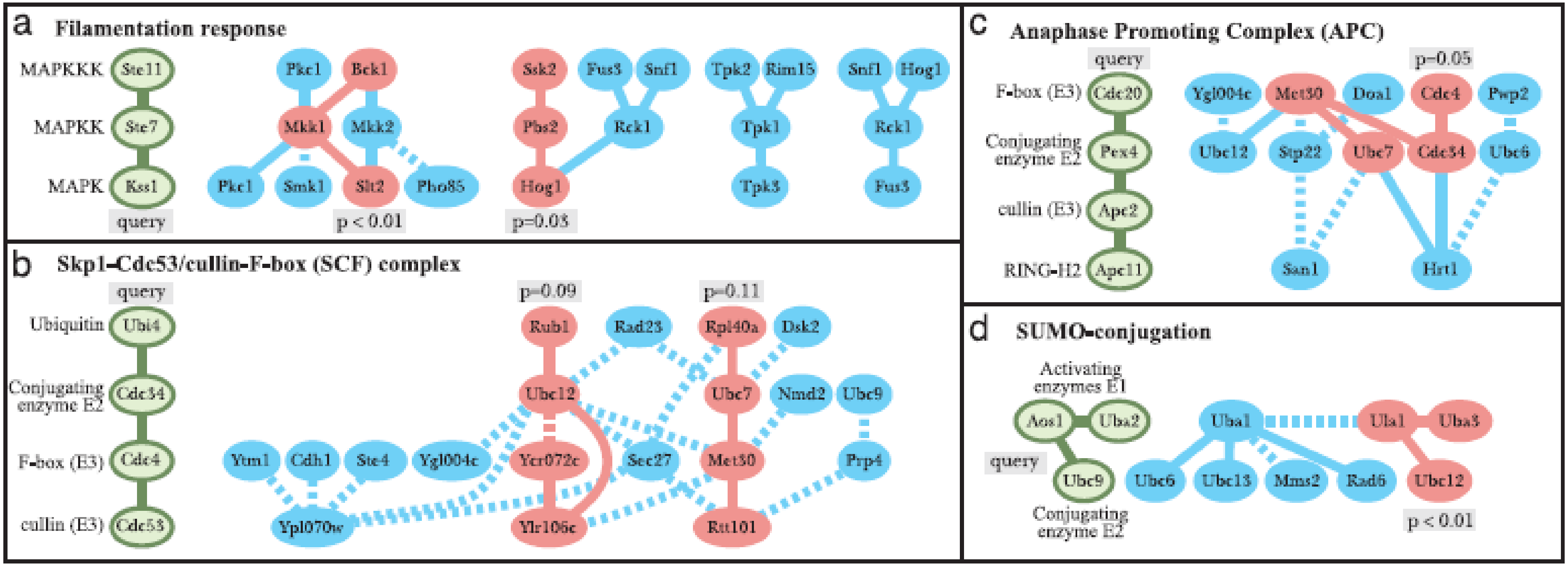
Color codes



Sample Alignment **Hyperlinked Complex**

```
Alignment 37   5.900
YHL007C (Ste20)   YJL095W (BCK1)   bypass requirement for protein kinase C
       |                *             homolog
YHR023W (Myo1)    YKL129C (MYO3)   myosin I
       |                *
YFL039C (Act1)    YJR065C (ARP3)   actin-related gene

Alignment 82   5.290
YHL007C (Ste20)  YFR014C (CMK1)  Calmodulin-dependent protein kinase
       *               *
YHR023W (Myo1)   YAL029C (MYO4)  Required for mother-specific HO expression, needed
       *               |          for the accumulation in daughter nuclei of Ash1p
YFL039C (Act1)   YDL029W (ARP2)  Involved in endocytosis and membrane growth and polarity
                                  accumulation in daughter nuclei of Ash1p
```

Yeast network



cross species importance

- Cross species comparisons
- Application:
  - study and treatment of disease
  - directing drugs to pathways that are present in a pathogenic organism but absent in human host
  - Denovo pathway

Thank You