Graemlin: General and robust alignment of multiple large interaction networks

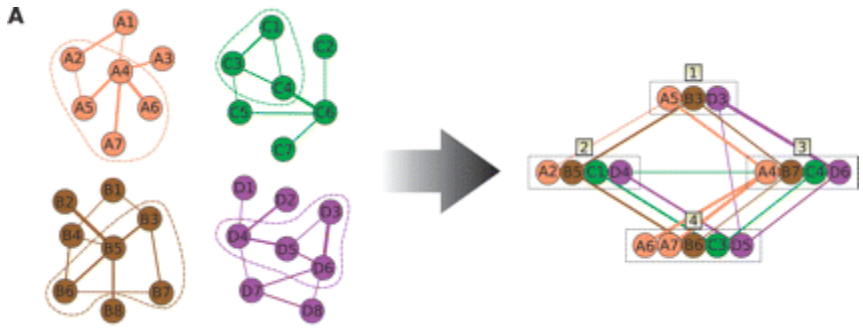
Flannick et al.

Introduction

- Graemlin: BLAST for ppi networks
  - Fast
  - Heuristic
  - Dual or multi alignments
- First quantitative comparison of network aligners

Methods

- What is a graph alignment algorithm?
  - Grouping proteins into equivalence classes
  - Determining an alignment's score
  - Choosing what to align



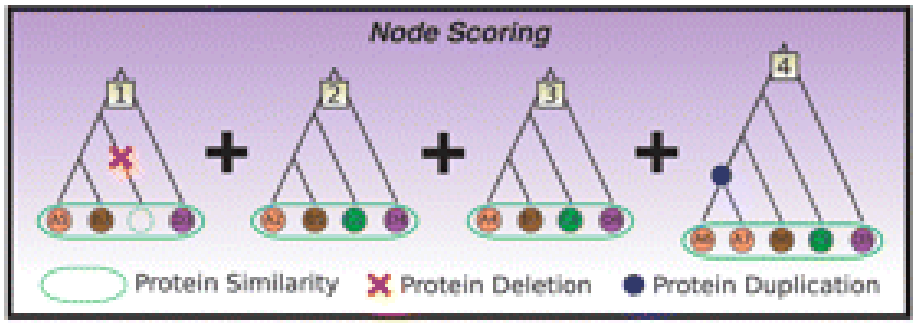
Methods: Scoring

- Log ratio of model versus background
- Score the equivalence class
- Score the edges

Methods: Node Scoring

- Model trained on pairs from COGs (Clusters of Orthologous Groups)
- Random pairs as background
- Score based on mutations/duplications/ insertions etc.
- Node score is sum-of-pairs

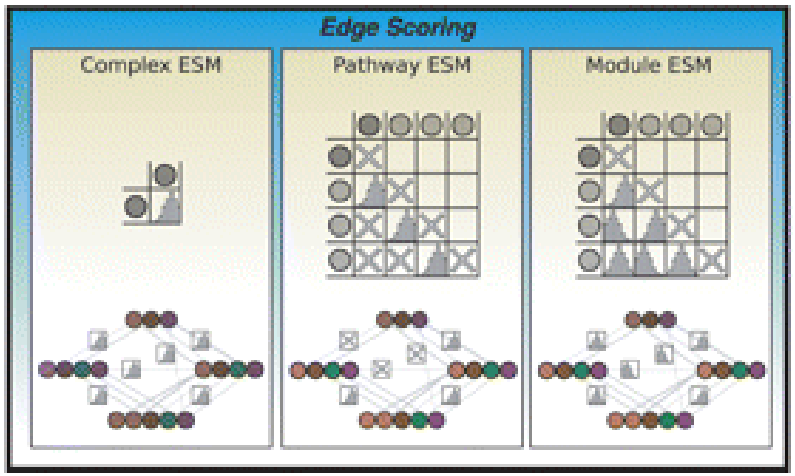
B



Methods: Edge Scoring

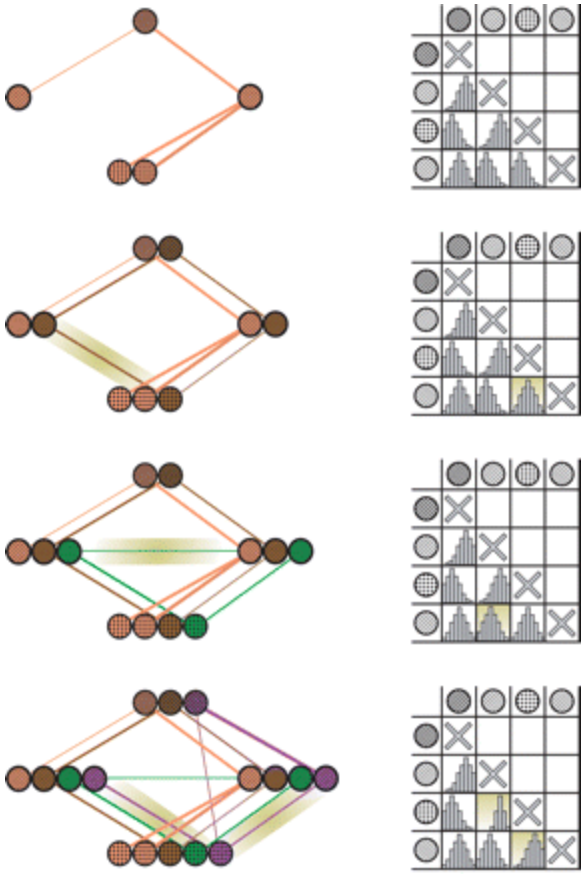
- Edge scoring matrix determines edge weights
- Different scoring methods prioritize different kinds of clusters

C



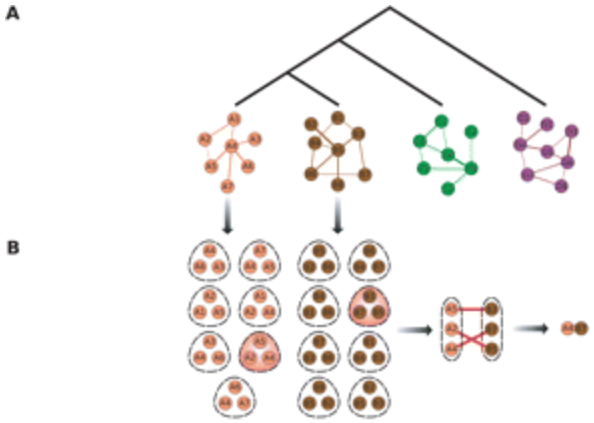
Methods: Edge Scoring

- Refines the matrix as more species are added



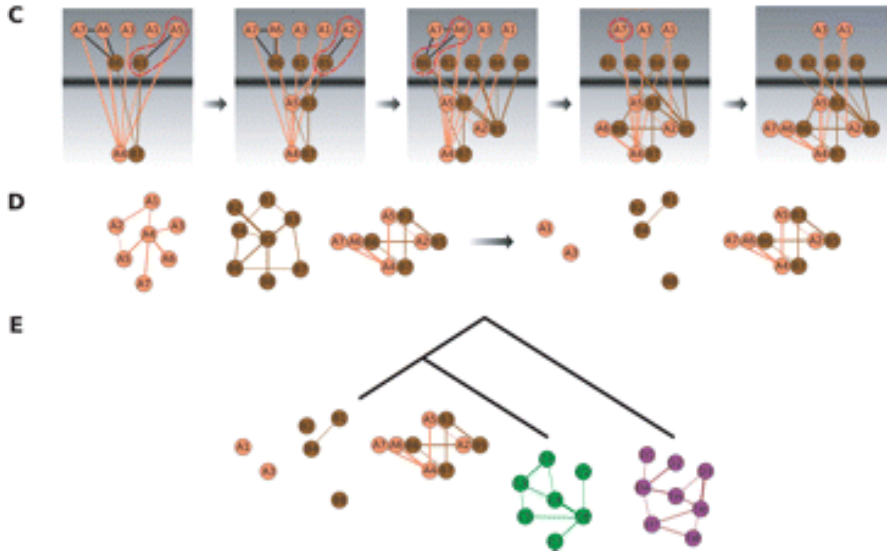
Methods: Alignment

- Pairwise
  - D-cluster = k-mer
    - In ONE species
    - Heuristic!
  - Join pairs of d-clusters over some threshold



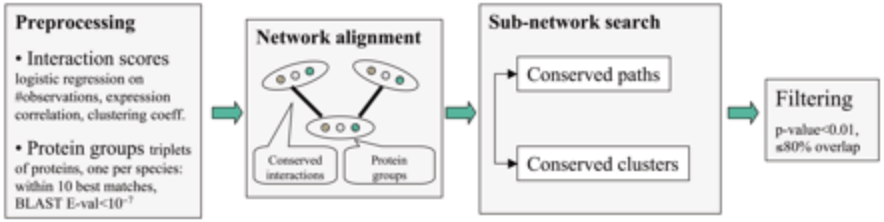
Methods: Alignment

- Expand the aligned d-cluster greedily
  - Stop when no further addition increases the score
- Multiply align by using our combined network as a single network



Results

- Compared to NetworkBLAST and MaWISh
  - NetworkBLAST

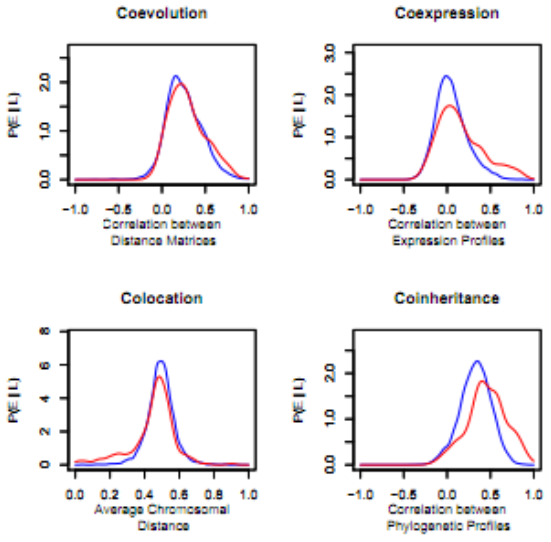


(Sharan et al 2004)

  - Mawish
    - Builds a fixed alignment graph, starts from nodes and greedily finds maximally scoring subgraphs

Results: Building a network with SRINI

- SRINI uses four factors
- Not from direct experimental data
- Recapitulates KEGGS in prokaryotes only



(b) Evidence vs. Training Set

(Srinivasan et al 2006)

Results

- KEGGs are manually annotated pathways
- Sensitivity: what is a hit?
  - ‘aligned at least three proteins in each species to their correct counterparts in all other species’
    - Sensitivity
- Also a specificity measure: percent of alignments that are GO-enriched

Results: Pairwise

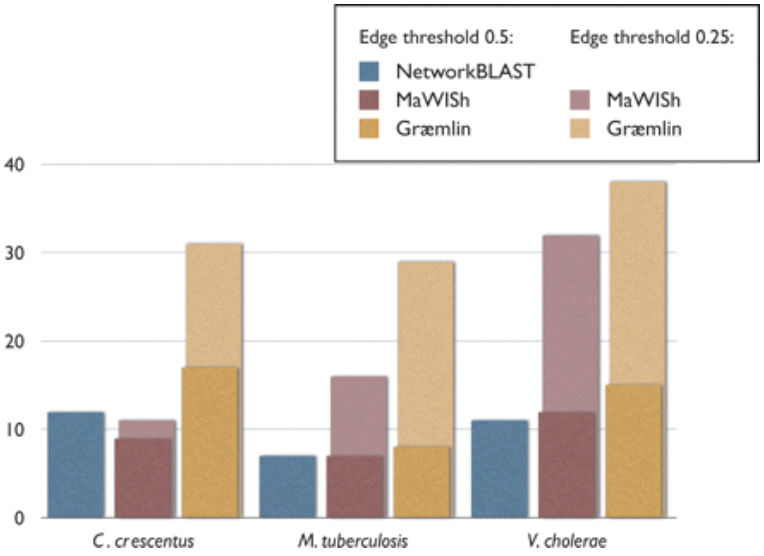


Table 3. Results on pairwise alignment of complete networks thresholded at 0.5

		KEGGs hit	KEGG coverage	Alignments enriched	Running time (sec)
E. coli vs. *C. crescentus*					
MaWISh		9 (20%)	32%	72%	3
NetworkBLAST	Pathway	6 (14%)	28%	61%	9624
	Complex	12 (27%)	49%	72%	
Græmlin	Pathway	15 (34%)	47%	68%	21
	Complex	17 (39%)	45%	67%	11
E. coli vs. *M. tuberculosis*					
MaWISh		7 (13%)	20%	85%	3
NetworkBLAST	Pathway	7 (13%)	24%	88%	301
	Complex	7 (13%)	32%	88%	
Græmlin	Pathway	8 (15%)	36%	89%	11
	Complex	8 (15%)	39%	89%	8
E. coli vs. *V. cholerae*					
MaWISh		12 (31%)	35%	64%	3
NetworkBLAST	Pathway	10 (26%)	35%	58%	8797
	Complex	11 (28%)	41%	64%	
Græmlin	Pathway	19 (49%)	48%	75%	13
	Complex	15 (38%)	55%	74%	12
E. coli vs. *S. coelicolor*					
MaWISh		N/A	N/A	N/A	N/A
NetworkBLAST	Pathway	6 (14%)	23%	46%	122,168
	Complex	10 (23%)	67%	95%	
Græmlin	Pathway	8 (19%)	58%	88%	734
	Complex	9 (21%)	59%	85%	829

For each pair of species, we performed complete network-to-network alignment using MaWISh and Græmlin. For each tested method, shown, from left, is the total number of KEGG pathways hit by an alignment, the fraction of KEGG pathways hit by an alignment, the average coverage of a KEGG pathway, the percentage of enriched alignments, and the total running time. We calculated the average coverage of KEGGs with respect to only those KEGGs that an aligner hit, and measured running time in CPU-seconds.

Results: Multiple

Table 4. Results on multiple alignment of complete networks

		KEGGs hit	KEGG coverage	Alignments enriched	Running time (sec)
0.25 threshold					
E. coli vs. *C. crescentus* vs. *V. cholerae*					
Græmlin	Pathway	27 (57%)	68%	72%	329
	Complex	29 (62%)	71%	79%	251
E. coli vs. *C. jejuni* vs. *H. pylori*					
Græmlin	Pathway	16 (57%)	57%	87%	44
	Complex	17 (61%)	63%	89%	43
0.5 threshold					
E. coli vs. *C. crescentus* vs. *V. cholerae*					
NetworkBLAST	Pathway Complex	N/A	N/A	N/A	$>10^6$
Græmlin	Pathway	7 (26%)	67%	72%	63
	Complex	9 (33%)	62%	75%	38
E. coli vs. *C. jejuni* vs. *H. pylori*					
NetworkBLAST	Pathway	5 (33%)	41%	94%	32,394
	Complex	4 (27%)	38%	96%	
Græmlin	Pathway	3 (20%)	74%	82%	12
	Complex	3 (20%)	72%	79%	12

We performed three-way multiple network alignment using NetworkBLAST and Græmlin; the columns in this table are analogous to those in Table 3.

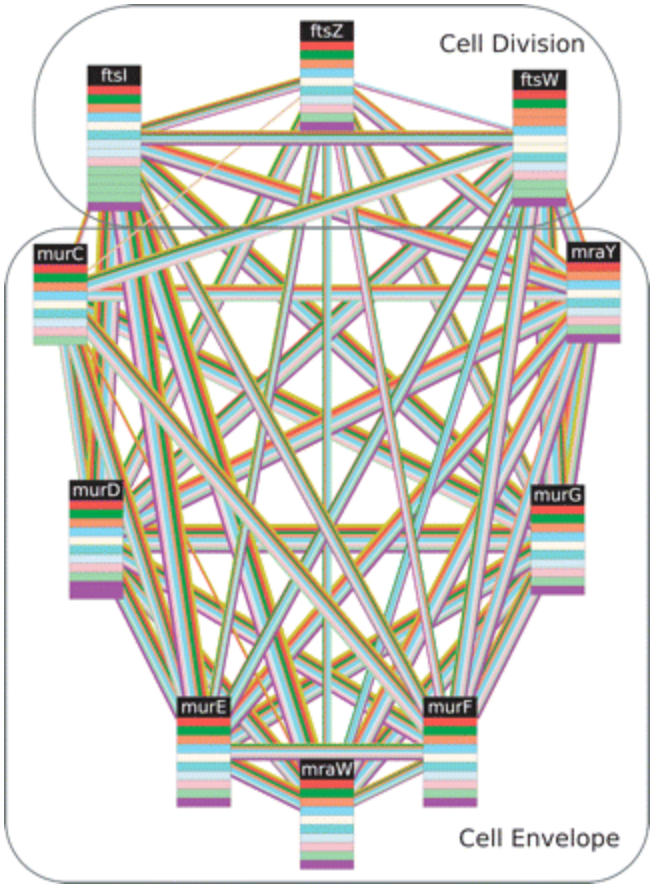
Results: Query

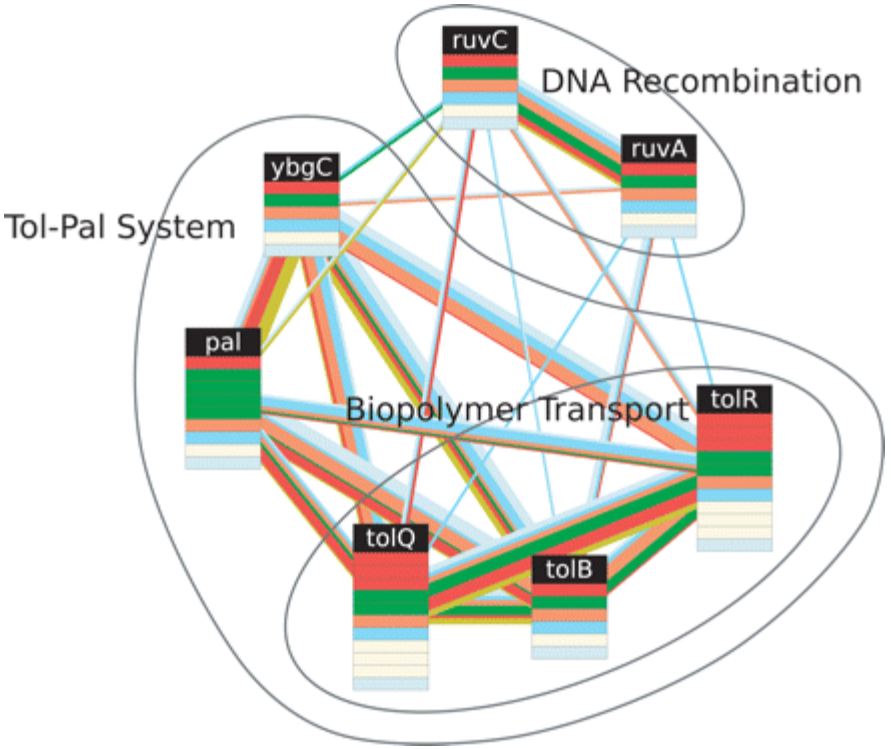
- Complex performs well?!
  - ‘most KEGGs that are alignable are highly connected, making the Complex ESM close to optimal’

Table 5. Results on alignment of a query network to a database thresholded at 0.5

		KEGGs hit	KEGG coverage	Running time (sec)
E. coli vs. *C. crescentus*				
MaWISh		15 (34%)	31%	37
NetworkBLAST	Pathway	8 (18%)	32%	3453
	Complex	10 (23%)	49%	
Græmlin	Pathway	20 (45%)	45%	17
	Complex	20 (45%)	47%	3
	Module	20 (45%)	48%	23
C. crescentus vs. *E. coli*				
MaWISh		9 (20%)	32%	130
NetworkBLAST	Pathway	10 (23%)	37%	4788
	Complex	10 (23%)	41%	
Græmlin	Pathway	15 (34%)	39%	6
	Complex	15 (34%)	42%	5
	Module	15 (34%)	42%	33
E. coli vs. *M. tuberculosis*				
MaWISh		10 (19%)	19%	93
NetworkBLAST	Pathway	12 (22%)	23%	3947
	Complex	12 (22%)	29%	
Græmlin	Pathway	17 (31%)	31%	3
	Complex	17 (31%)	35%	3

Results: Biological discovery





Conclusion

- Built a network alignment algorithm and a repeatable set of assessments
- Demonstrated ability to recover alignments from prokaryotes
- Builds on the BLAST model of alignment in a new context
- Fast; greedy; scales well