Gene Function Prediction

T. M. Murali

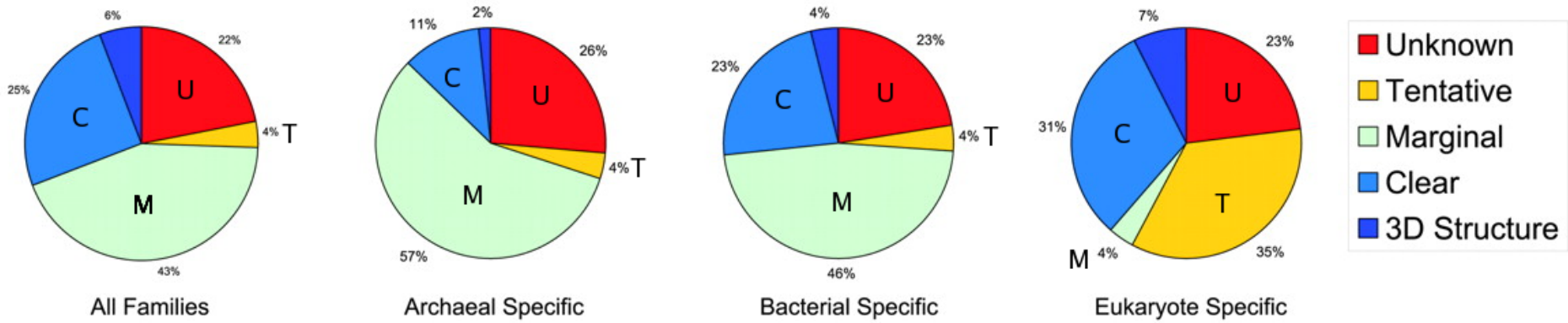
August 21, 2008

Data, Data, Data

- $\geq$ 150 genomes sequenced, 100 microbial and 50 eukaryotic.
- Computational identification of genes.
- Systematic gene knockouts.
- Gene expression data, proteomic data, metabolic data.
- Molecular interaction networks, metabolic pathways.

Roadblock: What do the Genes do?

"During the last few years, we have seen enormous strides in our abilities to sequence genomes, . . . With more than 150 complete genome sequences now available and many laboratories rushing into microarray analysis, proteomic initiatives, and even systems biology, it seems an appropriate time to consider not just the opportunities those sequences present, but also their shortcomings. By far the most serious problem is the quality and degree of completeness of the annotation of those genomes." (*Identifying Protein Function—A Call for Community Action*. Roberts RJ (2004), PLoS Biol 2(3): e42.)



Solution: Automated Functional Annotation

- Develop computational techniques that automatically integrate diverse source of data to predict function.
- Provide measures of confidence and statistical significance for each prediction.
- Present the predictions in a user-friendly manner to a biologist for designing experiments to validate prediction.

How do you Predict Function?

- Genes with similar sequences in different organisms are likely to have the same function.
- Use algorithms for computing sequence and structural similarity.
- Transfer the known function of a well-studied gene to a gene with a similar sequence that has no known functions.

How do you Predict Function?

- Genes with similar sequences in different organisms are likely to have the same function.
- Use algorithms for computing sequence and structural similarity.
- Transfer the known function of a well-studied gene to a gene with a similar sequence that has no known functions.

BUT

- 25% of the genes have no known sequence or structural similarity to any gene in any other organism (60% in *Plasmodium falciparum*).
- An additional 50% have poor annotations.

How do you Predict Function?

- Genes with similar sequences in different organisms are likely to have the same function.
- Use algorithms for computing sequence and structural similarity.
- Transfer the known function of a well-studied gene to a gene with a similar sequence that has no known functions.

BUT

- 25% of the genes have no known sequence or structural similarity to any gene in any other organism (60% in *Plasmodium falciparum*).
- An additional 50% have poor annotations.

We need techniques for functional annotation that go beyond sequence similarity.

What is Gene Function?

- Not an easy question to answer!
- A gene's function has many aspects.
- Different aspects are interesting to different biologists.
- There are many ways to describe a gene's function.
- Different groups of biologists have derived different vocabularies.
- A number of different functional catalogues exist: MultiFun (for *E. coli*), MIPS FunCat, structure-based (e.g., PFam/ProSite domains, SCOP), COG, EC, Uniprot

The Gene Ontology

- Collaborative effort to define a controlled vocabulary to describe gene and gene product attributes in any organism.

The Gene Ontology

- Collaborative effort to define a controlled vocabulary to describe gene and gene product attributes in any organism.
- Visit http://www.geneontology.org

The Gene Ontology

- Collaborative effort to define a controlled vocabulary to describe gene and gene product attributes in any organism.
- Visit http://www.geneontology.org
- Three Gene Ontology (GO) categories:

The Gene Ontology

- Collaborative effort to define a controlled vocabulary to describe gene and gene product attributes in any organism.
- Visit http://www.geneontology.org
- Three Gene Ontology (GO) categories: A gene product has
  - a *molecular function*: an activity, such as catalytic or binding activity, carried out by the gene product at the molecular level;

The Gene Ontology

- Collaborative effort to define a controlled vocabulary to describe gene and gene product attributes in any organism.
- Visit http://www.geneontology.org
- Three Gene Ontology (GO) categories: A gene product has
  - a *molecular function*: an activity, such as catalytic or binding activity, carried out by the gene product at the molecular level;
  - is used in a *biological process*: a series of events accomplished by one or more ordered assemblies of molecular functions; and

The Gene Ontology

- Collaborative effort to define a controlled vocabulary to describe gene and gene product attributes in any organism.
- Visit http://www.geneontology.org
- Three Gene Ontology (GO) categories: A gene product has
  - a *molecular function*: an activity, such as catalytic or binding activity, carried out by the gene product at the molecular level;
  - is used in a *biological process*: a series of events accomplished by one or more ordered assemblies of molecular functions; and
  - might be associated with a *cellular component*: a component of a cell that is part of some larger object, which may be an anatomical structure or a gene product group.

The Gene Ontology

- Collaborative effort to define a controlled vocabulary to describe gene and gene product attributes in any organism.
- Visit http://www.geneontology.org
- Three Gene Ontology (GO) categories: A gene product has
  - a *molecular function*: an activity, such as catalytic or binding activity, carried out by the gene product at the molecular level;
  - is used in a *biological process*: a series of events accomplished by one or more ordered assemblies of molecular functions; and
  - might be associated with a *cellular component*: a component of a cell that is part of some larger object, which may be an anatomical structure or a gene product group.
- For example, the gene product *cytochrome c* can be described by
  - the molecular function term *oxidoreductase activity*,
  - the biological process terms *oxidative phosphorylation* and *induction of cell death*, and
  - the cellular component terms *mitochondrial matrix* and *mitochondrial inner membrane*.

Jump to FLNs

Features of GO

- GO functions, processes, and components are described at multiple levels of detail.
- Explicit parent-child relationships between terms.

Features of GO

- GO functions, processes, and components are described at multiple levels of detail.
- Explicit parent-child relationships between terms. The terms form a directed acyclic graph (DAG).

Features of GO

- GO functions, processes, and components are described at multiple levels of detail.
- Explicit parent-child relationships between terms. The terms form a directed acyclic graph (DAG).
- A team of experts define GO terms.
- Annotations typically done by individual genome databases.

Features of GO

- GO functions, processes, and components are described at multiple levels of detail.
- Explicit parent-child relationships between terms. The terms form a directed acyclic graph (DAG).
- A team of experts define GO terms.
- Annotations typically done by individual genome databases.
- Evidence code attached to annotation:

Features of GO

- GO functions, processes, and components are described at multiple levels of detail.
- Explicit parent-child relationships between terms. The terms form a directed acyclic graph (DAG).
- A team of experts define GO terms.
- Annotations typically done by individual genome databases.
- Evidence code attached to annotation:
  - IDA: inferred from direct assay (enzyme assay, cell fractionation)
  - IPI: inferred from physical interaction (2-hybrid)
  - IGI: inferred from genetic interaction (suppressor, synthetic lethal)
  - IEP: inferred from expression pattern (microarray)
  - IMP: inferred from mutant phenotype
  - TAS: traceable author statement
  - IC: inferred by curator
  - NAS: non-traceable author statement

Features of GO

- GO functions, processes, and components are described at multiple levels of detail.
- Explicit parent-child relationships between terms. The terms form a directed acyclic graph (DAG).
- A team of experts define GO terms.
- Annotations typically done by individual genome databases.
- Evidence code attached to annotation:
  - IDA: inferred from direct assay (enzyme assay, cell fractionation)
  - IPI: inferred from physical interaction (2-hybrid)
  - IGI: inferred from genetic interaction (suppressor, synthetic lethal)
  - IEP: inferred from expression pattern (microarray)
  - IMP: inferred from mutant phenotype
  - TAS: traceable author statement
  - IC: inferred by curator
  - NAS: non-traceable author statement
  - ISS: inferred from sequence or structure similarity
  - IEA: inferred from electronic annotation

Features of GO

- GO functions, processes, and components are described at multiple levels of detail.
- Explicit parent-child relationships between terms. The terms form a directed acyclic graph (DAG).
- A team of experts define GO terms.
- Annotations typically done by individual genome databases.
- Evidence code attached to annotation:
  - IDA: inferred from direct assay (enzyme assay, cell fractionation)
  - IPI: inferred from physical interaction (2-hybrid)
  - IGI: inferred from genetic interaction (suppressor, synthetic lethal)
  - IEP: inferred from expression pattern (microarray)
  - IMP: inferred from mutant phenotype
  - TAS: traceable author statement
  - IC: inferred by curator
  - NAS: non-traceable author statement
  - ISS: inferred from sequence or structure similarity
  - IEA: inferred from electronic annotation
  - RCA: inferred from reviewed computational analysis

Potential Advantages of GO

- The vocabulary is controlled $\Rightarrow$ common vocabulary for all biologists.
- Designed to apply across species.
- Computed mappings from other functional catalgues to GO.
- The GO terms are constantly updated (actually a headache for functional annotation algorithms).
- Freely available to the community.

Moving Beyond GO

- GO does not describe many aspects of a gene's function:

Moving Beyond GO

- GO does not describe many aspects of a gene's function: which cells or tissues it is expressed in, which developmental stages it is expressed in, or its involvement in disease.

Moving Beyond GO

- GO does not describe many aspects of a gene's function: which cells or tissues it is expressed in, which developmental stages it is expressed in, or its involvement in disease.
- Other ontologies are being developed to meet these needs.
  - Open Biomedical Ontologies: http://obo.sourceforge.net/
  - Ontology Working Group of the Microarray Gene Expression Data Society (MGED): http://mged.sourceforge.net/ontologies/OntologyResources.php

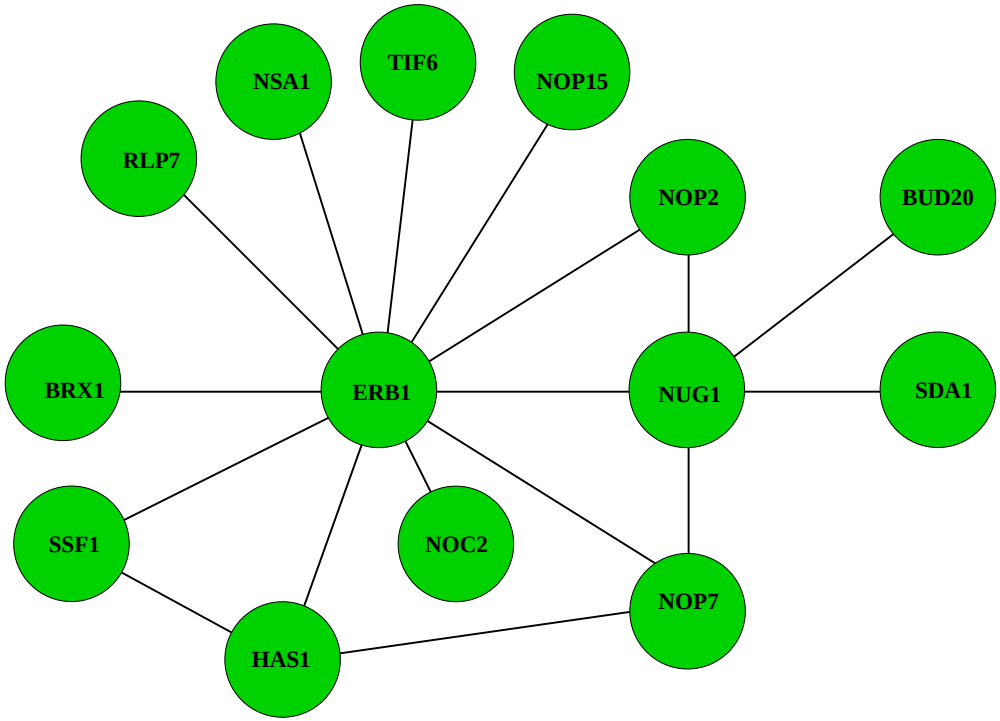
Moving Beyond GO

- GO does not describe many aspects of a gene's function: which cells or tissues it is expressed in, which developmental stages it is expressed in, or its involvement in disease.
- Other ontologies are being developed to meet these needs.
  - Open Biomedical Ontologies: http://obo.sourceforge.net/
  - Ontology Working Group of the Microarray Gene Expression Data Society (MGED): http://mged.sourceforge.net/ontologies/OntologyResources.php
- "Cross-products" of different ontologies: combine different (independent) ontologies to derive richer vocabularies.

Moving Beyond GO

- GO does not describe many aspects of a gene's function: which cells or tissues it is expressed in, which developmental stages it is expressed in, or its involvement in disease.
- Other ontologies are being developed to meet these needs.
  - Open Biomedical Ontologies: http://obo.sourceforge.net/
  - Ontology Working Group of the Microarray Gene Expression Data Society (MGED): http://mged.sourceforge.net/ontologies/OntologyResources.php
- "Cross-products" of different ontologies: combine different (independent) ontologies to derive richer vocabularies.
- "For example, by combining the developmental terms in the GO process ontology with a second ontology that describes Drosophila anatomical structures, we could create an ontology of fly development."
- "We could create an ontology of biosynthetic pathways by combining the biosynthesis terms in the GO process ontology with a chemical ontology."

Functional Linkage Networks



- A *functional linkage network* (FLN) is a graph where each node corresponds to a gene and each edge connects two genes that may share a similar function.
- An edge may not indicate which function the connected genes share.

Constructing FLNs

Constructing FLNs

- Organism specific
- Cross-organism

Constructing FLNs

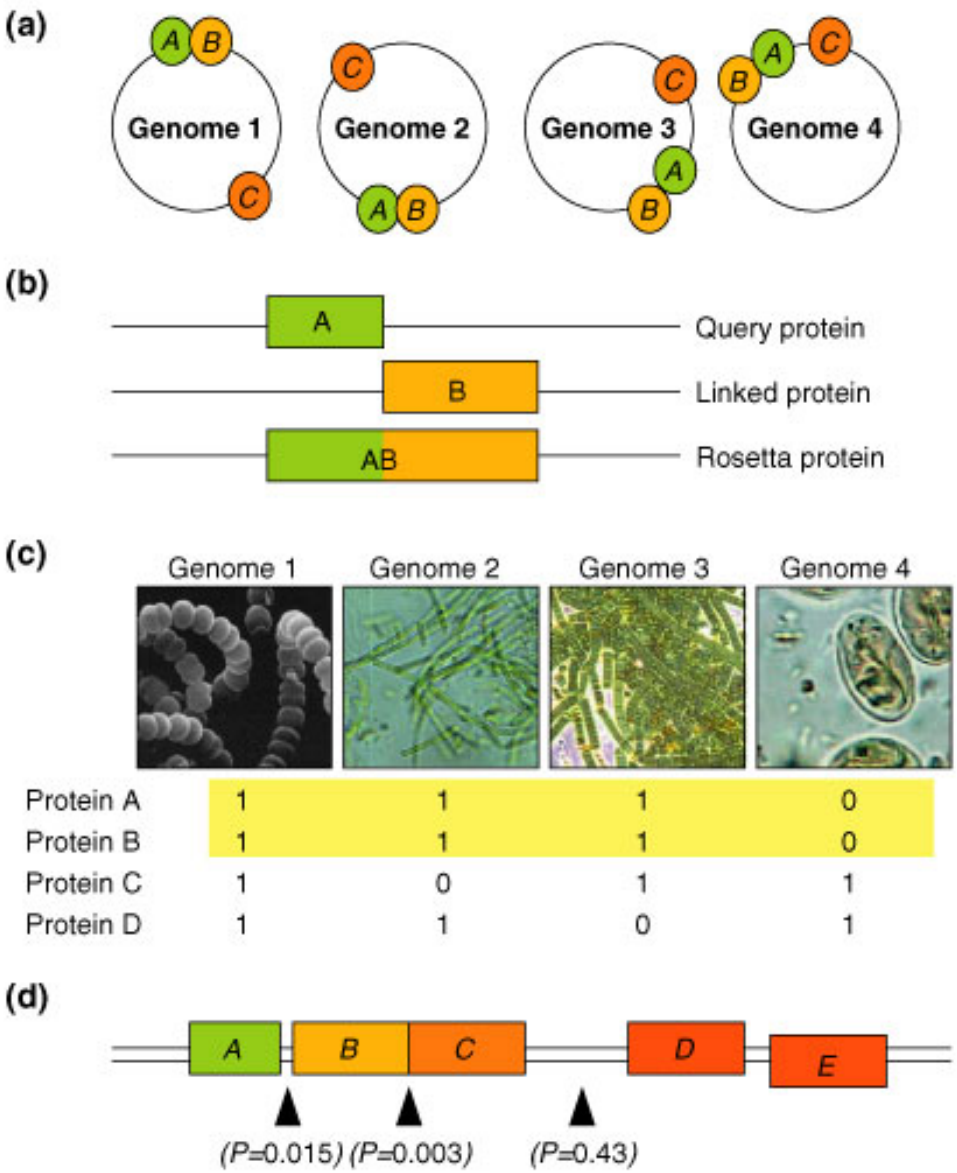
- Organism specific
  - Co-expression from DNA microarray data.
  - Protein products interact.
  - Genes co-regulated by the same transcription factor.
  - Double mutants are lethal (synthetic lethality).
  - Knockout mutants have the same metabolic profiles.
- Cross-organism

Constructing FLNs

- Organism specific
  - Co-expression from DNA microarray data.
  - Protein products interact.
  - Genes co-regulated by the same transcription factor.
  - Double mutants are lethal (synthetic lethality).
  - Knockout mutants have the same metabolic profiles.
- Cross-organism
  - Information on co-evolution encoded in genomic context.

▸ Onward to Challenges

Cross-Organism Functional Associations



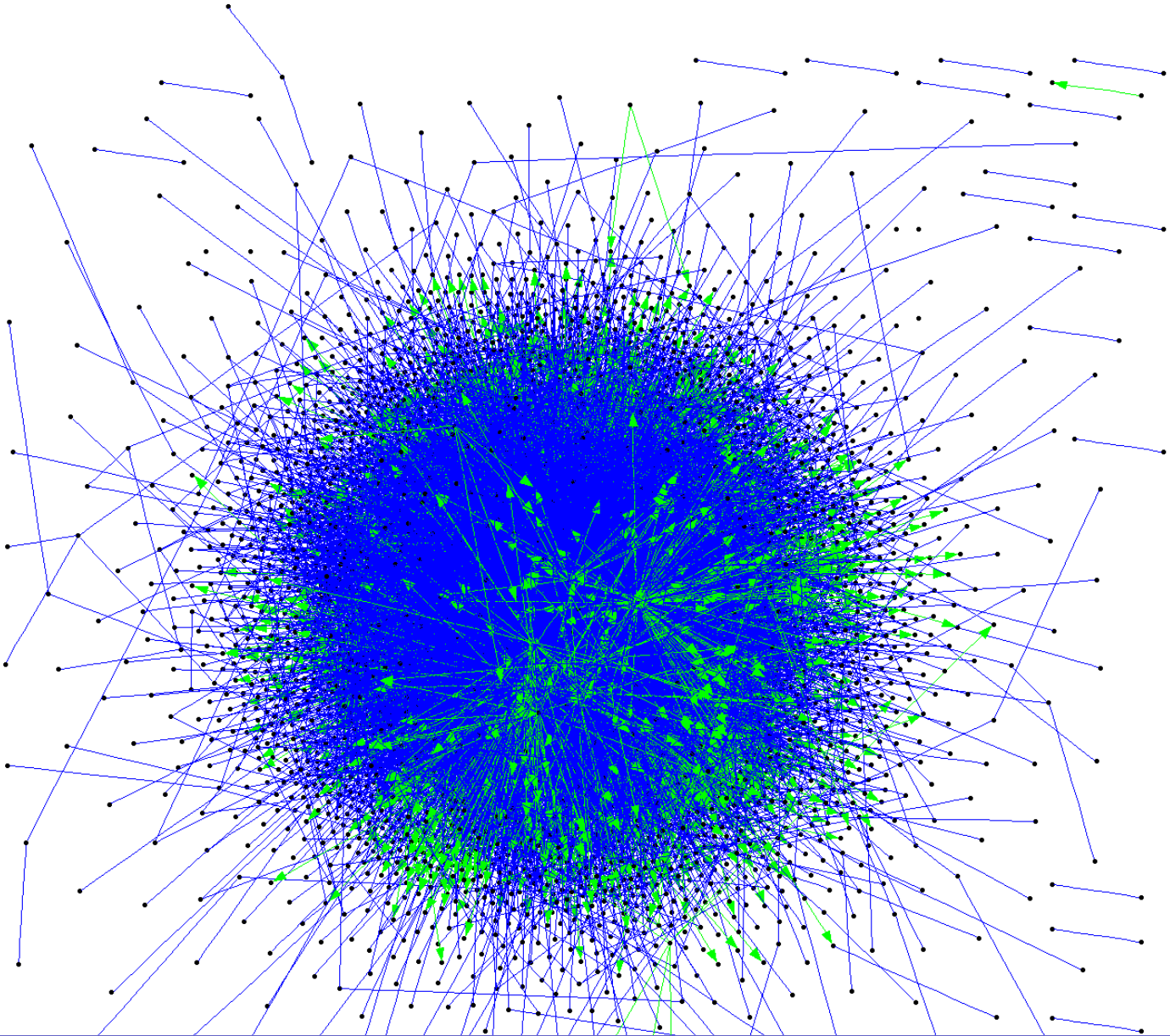
Previous Research on Functional Links

- Databases: BIND, DIP, GRID, IDSERVE, PROLINKS, PREDICTOME, REACTOME, STRING, ...
- Techniques for predicting functional associations, e.g., protein-protein interactions (Jansen et al., Science, 302, 2003; Zhang et al., BMC Bioinformatics, 5, 2005).
- Techniques for integrating diverse pieces of evidence into a single integrated FLN (Lee et al., Science, 306, 2005).

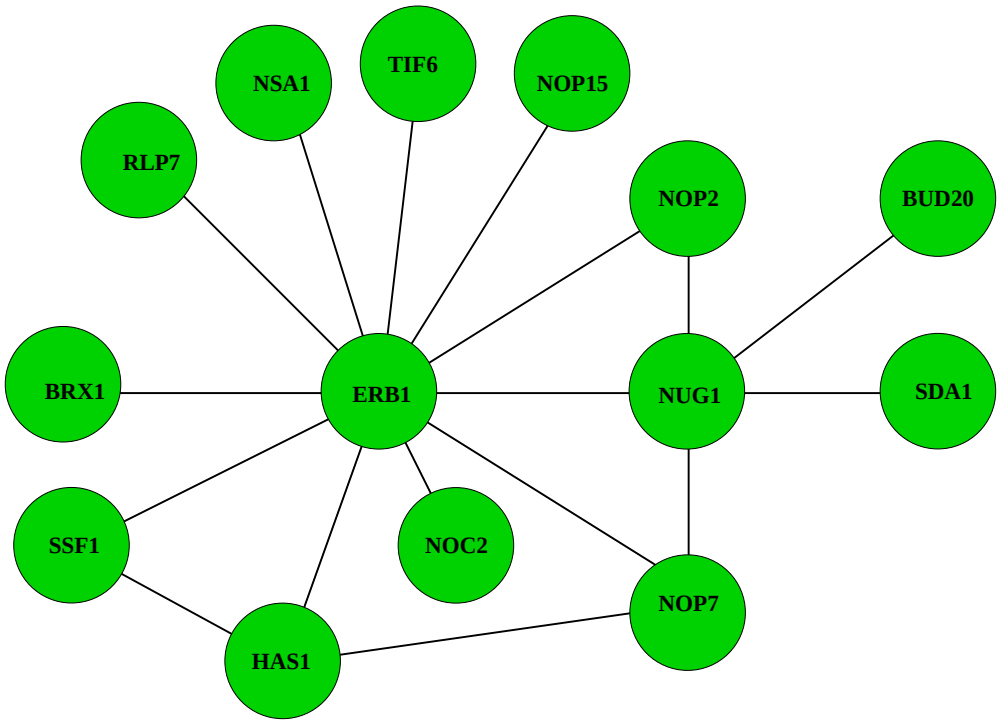
Previous Research on Functional Links

- Databases: BIND, DIP, GRID, IDSERVE, PROLINKS, PREDICTOME, REACTOME, STRING, ...
- Techniques for predicting functional associations, e.g., protein-protein interactions (Jansen et al., Science, 302, 2003; Zhang et al., BMC Bioinformatics, 5, 2005).
- Techniques for integrating diverse pieces of evidence into a single integrated FLN (Lee et al., Science, 306, 2005).
- How do we systematically use FLNs to make robust and quantified predictions of function?

Example of an FLN in Saccharomyces cerevisiae

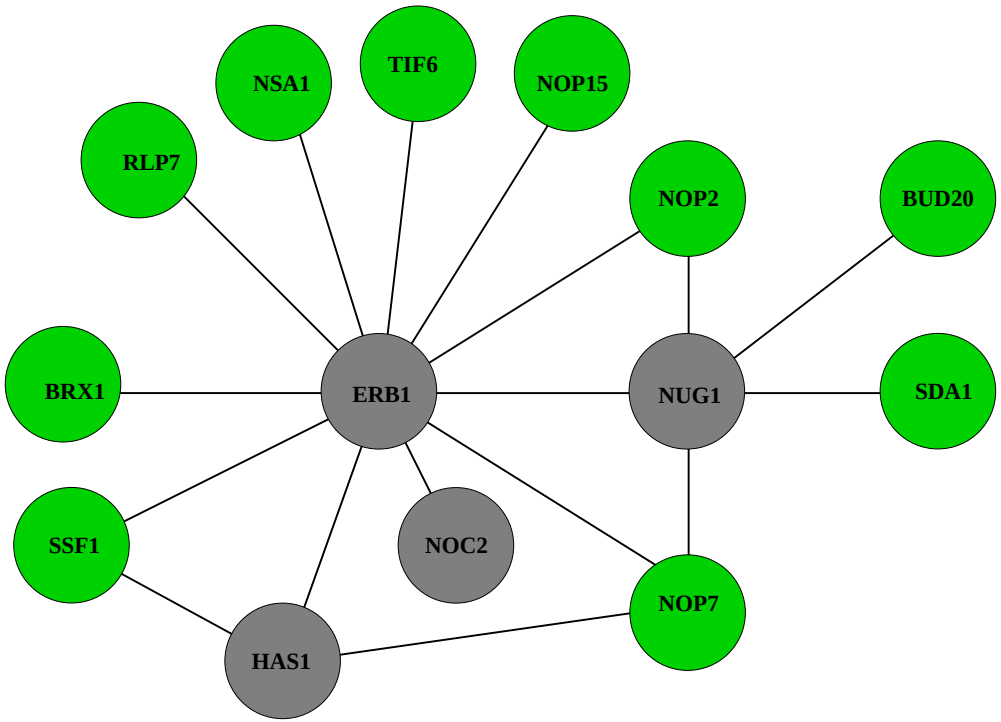


Why is Functional Annotation Difficult?



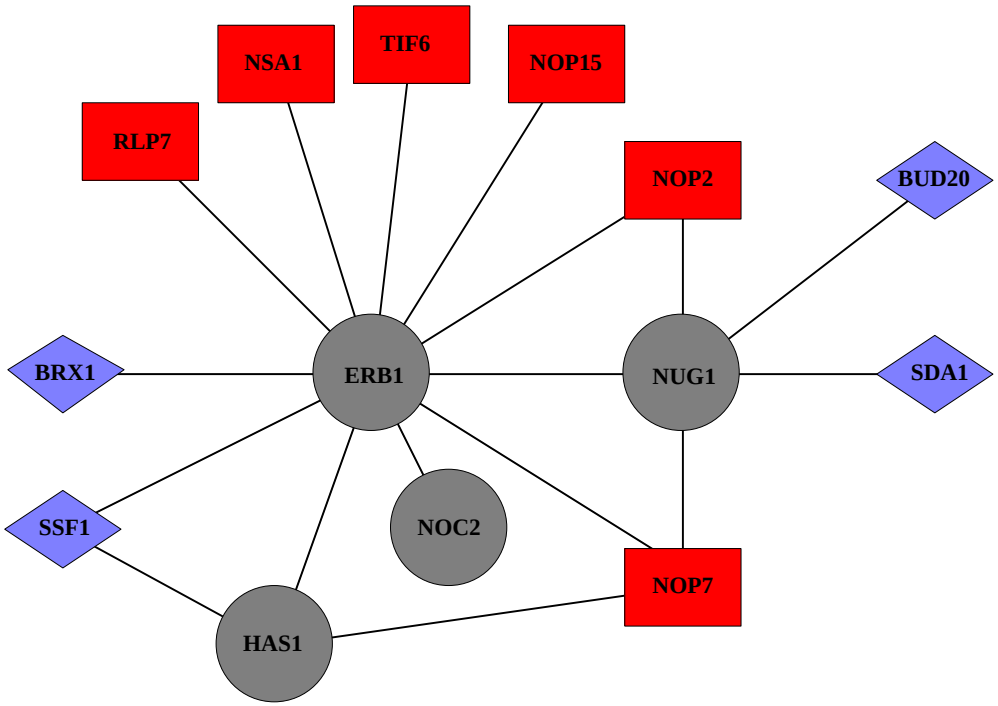
- Functional associations are not perfect indicators of shared function.

Why is Functional Annotation Difficult?



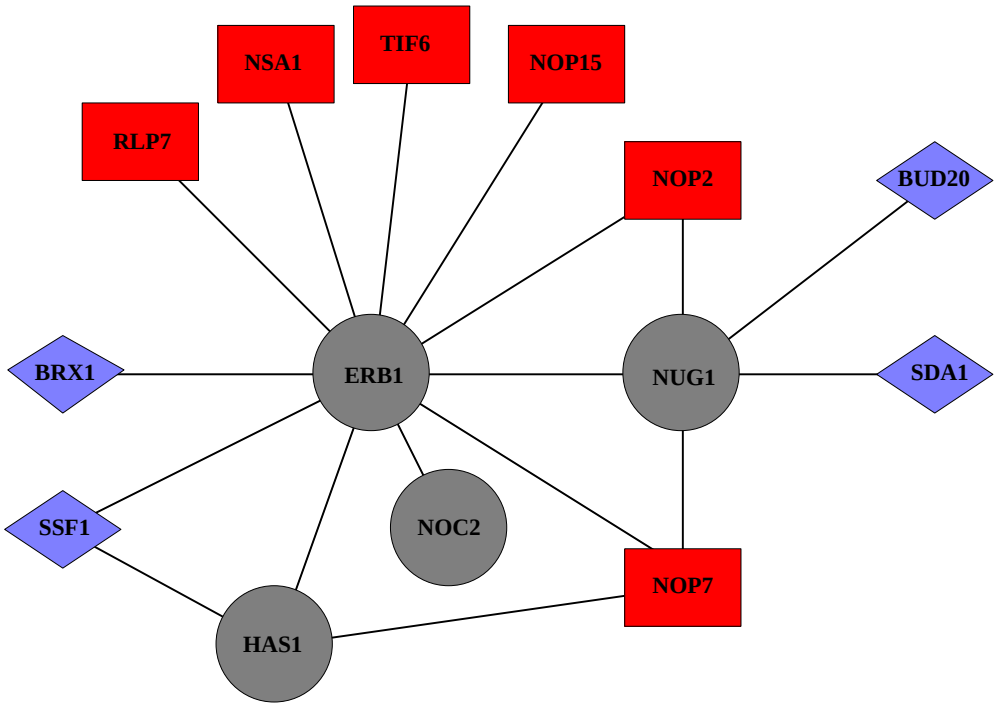
- Functional associations are not perfect indicators of shared function.
- 20–30% of genes of unknown function have only such genes as neighbours.

Why is Functional Annotation Difficult?



- Functional associations are not perfect indicators of shared function.
- 20–30% of genes of unknown function have only such genes as neighbours.
- Neighbourhood structure is ambiguous.

The GAIN System

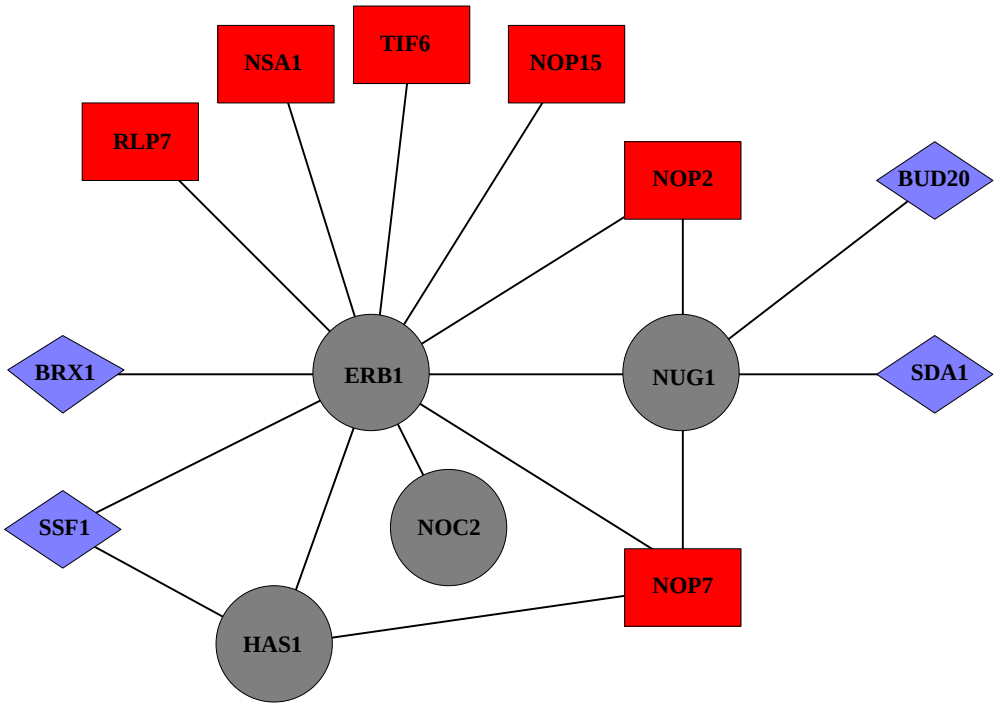


Gene Annotation Using Integrated Networks (GAIN):

- Propagate evidence systematically across the entire FLN.
- Integrate information from different sources to improve robustness.

(Karaoz, Murali, Letovsky, Zheng, Ding, Cantor and Kasif, "Whole genome annotation using evidence integration in functional linkage networks," *PNAS*, 2004, 101, 2888–2893.)

The GAIN System



Gene Annotation Using Integrated Networks (GAIN):

- Propagate evidence systematically across the entire FLN.
- Integrate information from different sources to improve robustness: protein-protein interactions and gene expression data.

(Karaoz, Murali, Letovsky, Zheng, Ding, Cantor and Kasif, "Whole genome annotation using evidence integration in functional linkage networks," *PNAS*, 2004, 101, 2888–2893.)

Overview of the GAIN Pipeline

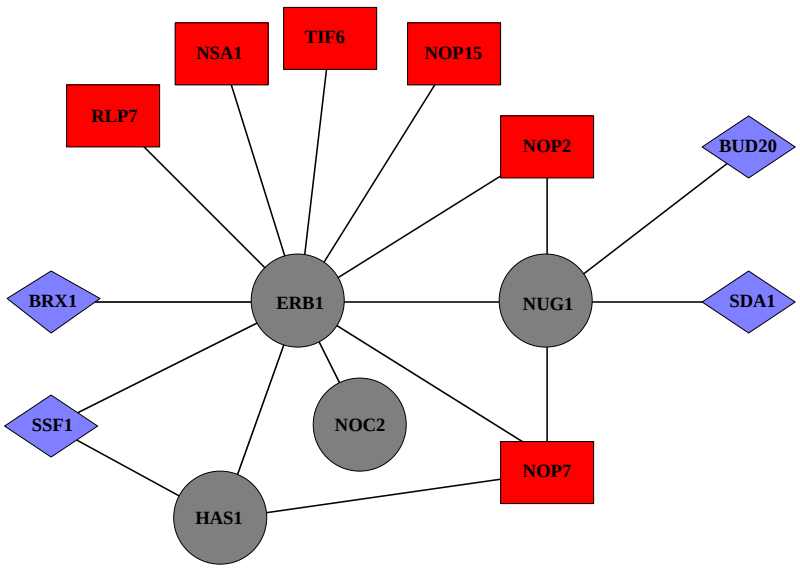
- Inputs: Functional genomic data sets, GO functional annotations.
- Outputs: For each function in GO, a set of genes predicted to have that function.

1. Construct FLN G from functional genomic data sets.
2. For each function f in GO
   - 2.1 Construct a labelled FLN G_f for f.
   - 2.2 Propagate the label f or $\mathrm{not}f$ across G_f.
   - 2.3 Output set of genes that have been assigned the function f.

- Can predict multiple functions for a gene.

Labelled FLNs

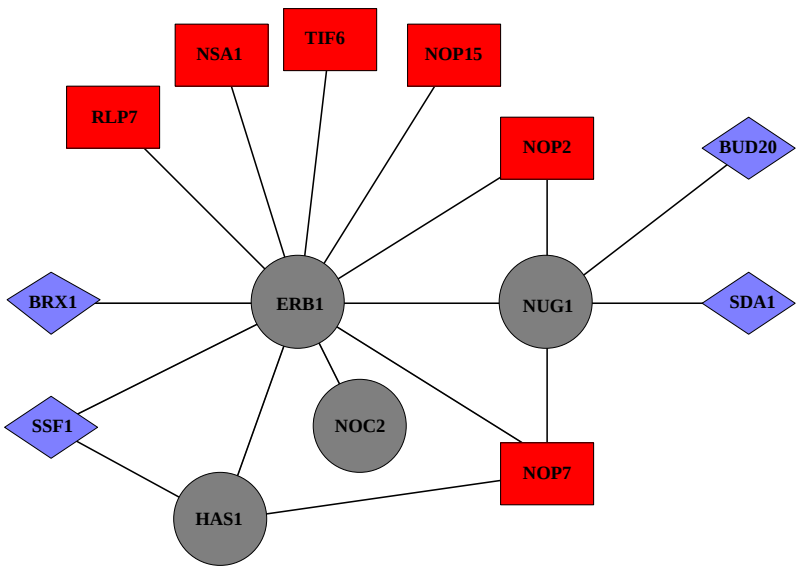
- *Labelled FLN* G_f for a function $f \equiv$ the FLN G with states (labels) attached to nodes.



- Each node i has an associated state s_i:
  - $s_i = 1$: gene i is annotated with f.
  - $s_i = -1$: gene i is annotated with another function f'.
  - $s_i = 0$: otherwise.
- An edge between nodes i and j has a weight w_{ij}.

Hopfield Networks

- Functional linkage graph $\longrightarrow$ discrete Hopfield network.
  - Gene $\equiv$ node.
  - Interaction $\equiv$ edge.



- Build a separate Hopfield network for each function.
- Given a function f, each node i has an associated state s_i:
  - $s_i = 1$: gene i is annotated with f.
  - $s_i = 0$: gene i is hypothetical.
  - $s_i = -1$: gene i is annotated with another function f'.
- An edge between nodes i and j has a weight w_{ij}.

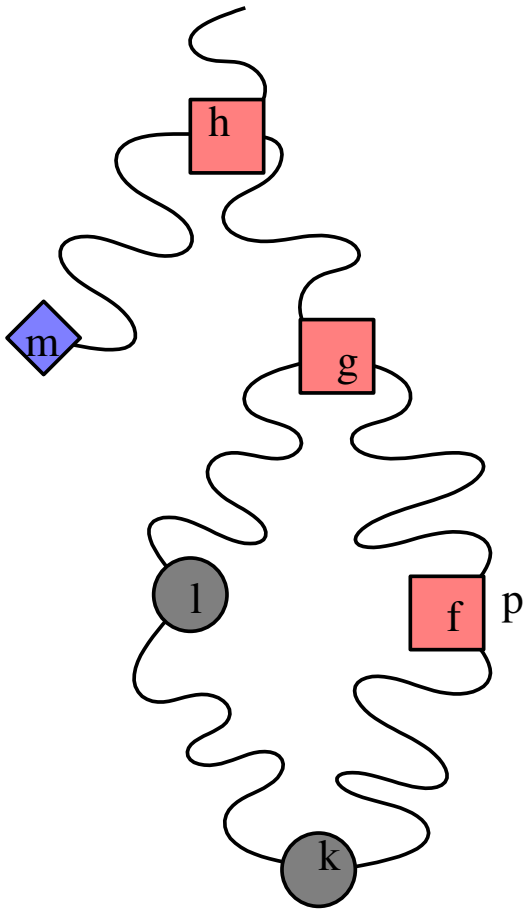
Skip Node States

Assigning Node States

- Assigning node states correctly is not a trivial manner.
- We must respect/exploit GO's hierarchical structure .

Assigning Node States

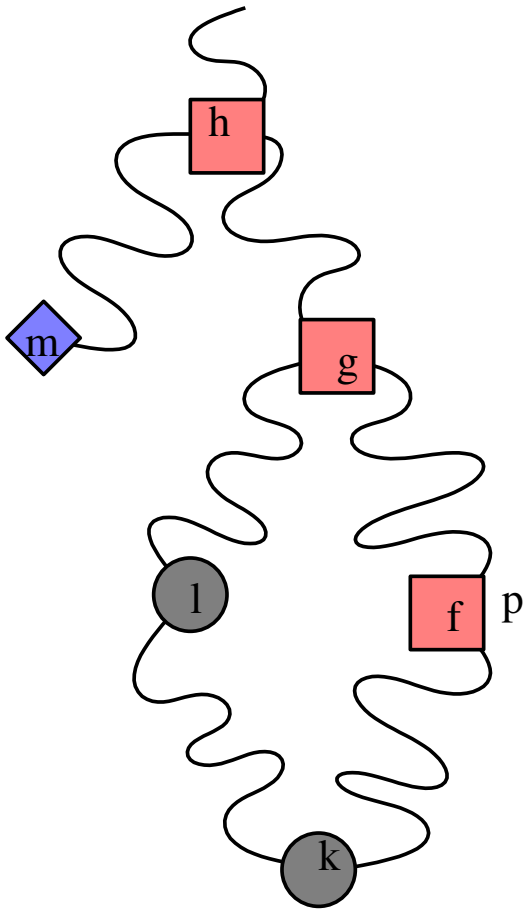
- Assigning node states correctly is not a trivial manner.
- We must respect/exploit GO's hierarchical structure .



- What is state of gene p with respect to function
  - f:
  - g:
  - h:
  - m:
  - k:
  - l:

Assigning Node States

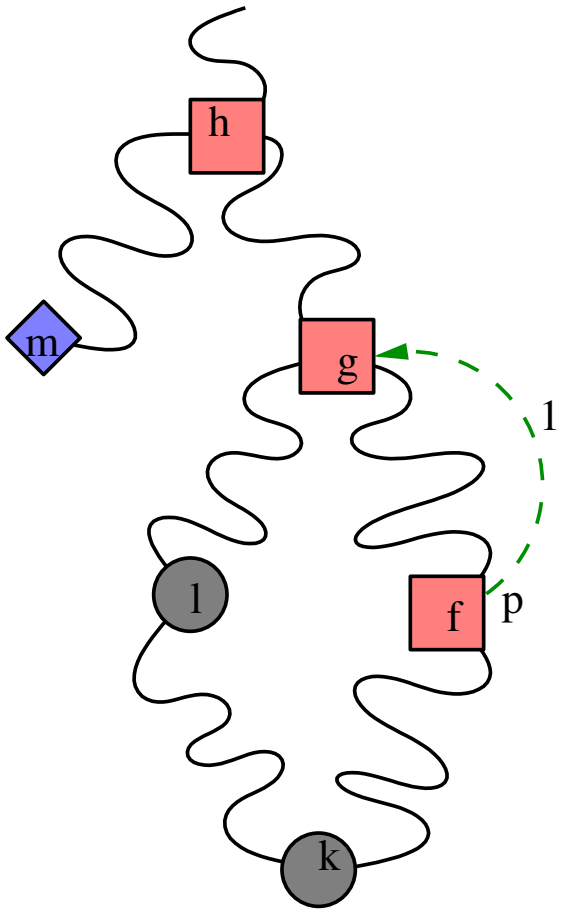
- Assigning node states correctly is not a trivial manner.
- We must respect/exploit GO's hierarchical structure .



- What is state of gene p with respect to function
  - f: 1
  - g:
  - h:
  - m:
  - k:
  - l:

Assigning Node States

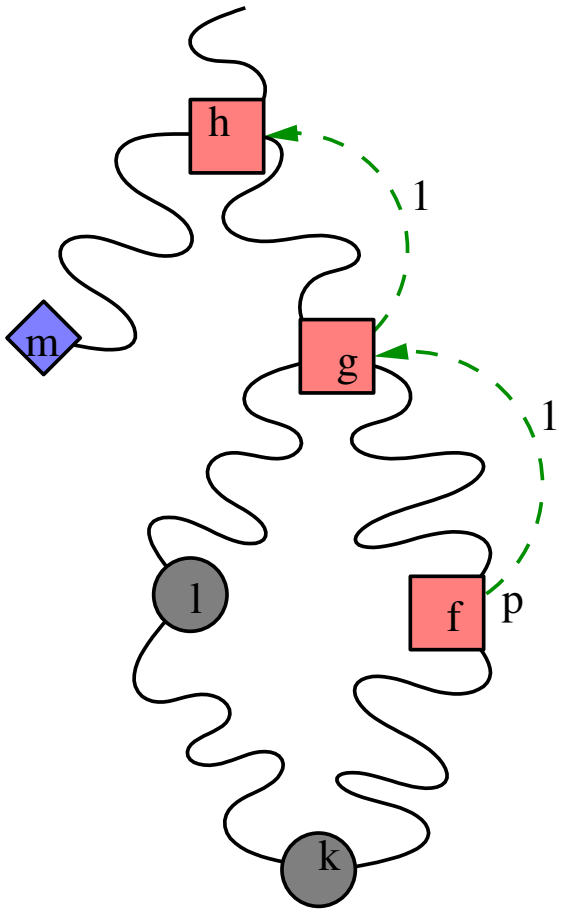
- Assigning node states correctly is not a trivial manner.
- We must respect/exploit GO's hierarchical structure .



- What is state of gene p with respect to function
  - f: 1
  - g: 1
  - h:
  - m:
  - k:
  - l:

Assigning Node States

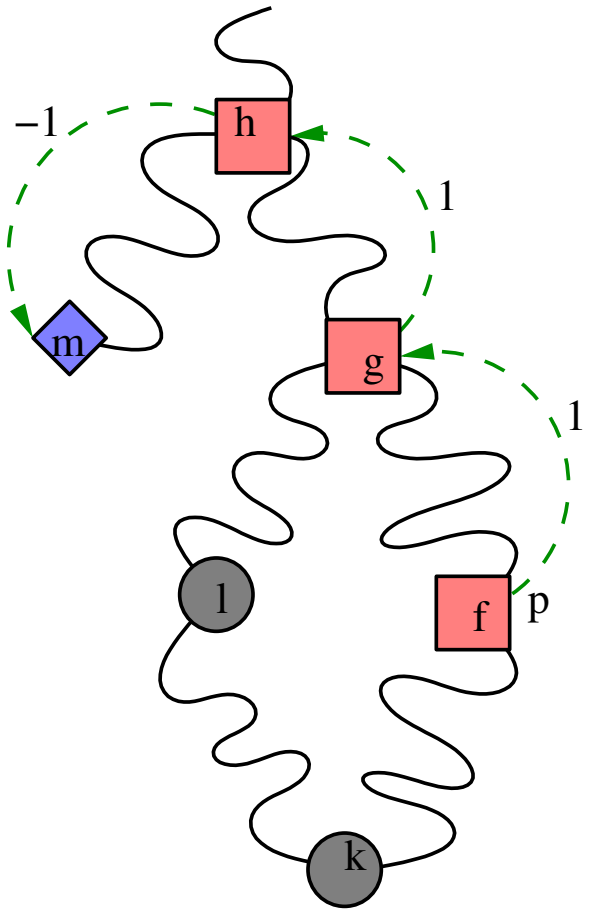
- Assigning node states correctly is not a trivial manner.
- We must respect/exploit GO's hierarchical structure .



- What is state of gene p with respect to function
  - f: 1
  - g: 1
  - h: 1
  - m:
  - k:
  - l:

Assigning Node States

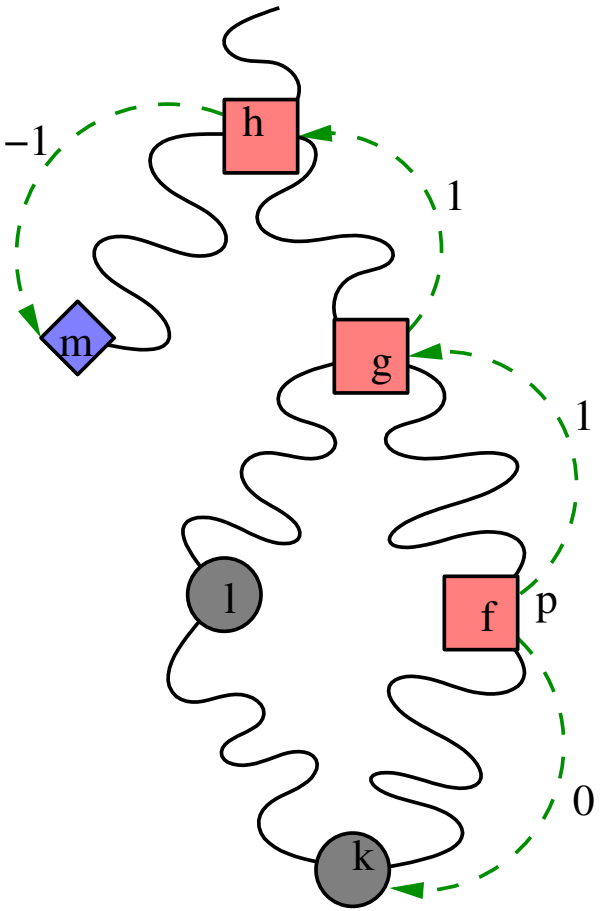
- Assigning node states correctly is not a trivial manner.
- We must respect/exploit GO's hierarchical structure .



- What is state of gene p with respect to function
  - f: 1
  - g: 1
  - h: 1
  - m: -1
  - k:
  - l:

Assigning Node States

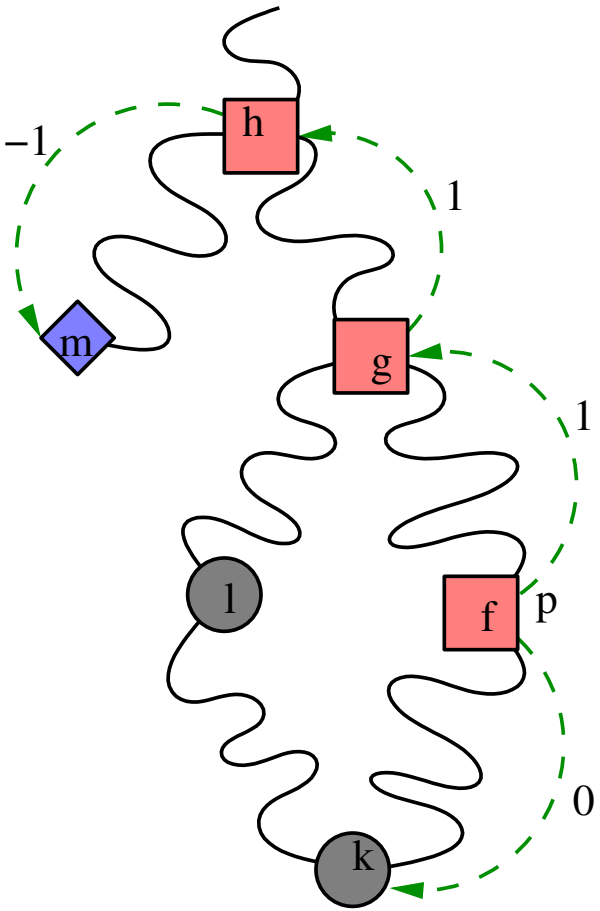
- Assigning node states correctly is not a trivial manner.
- We must respect/exploit GO's hierarchical structure .



- What is state of gene p with respect to function
  - f: 1
  - g: 1
  - h: 1
  - m: -1
  - k: 0
  - l:

Assigning Node States

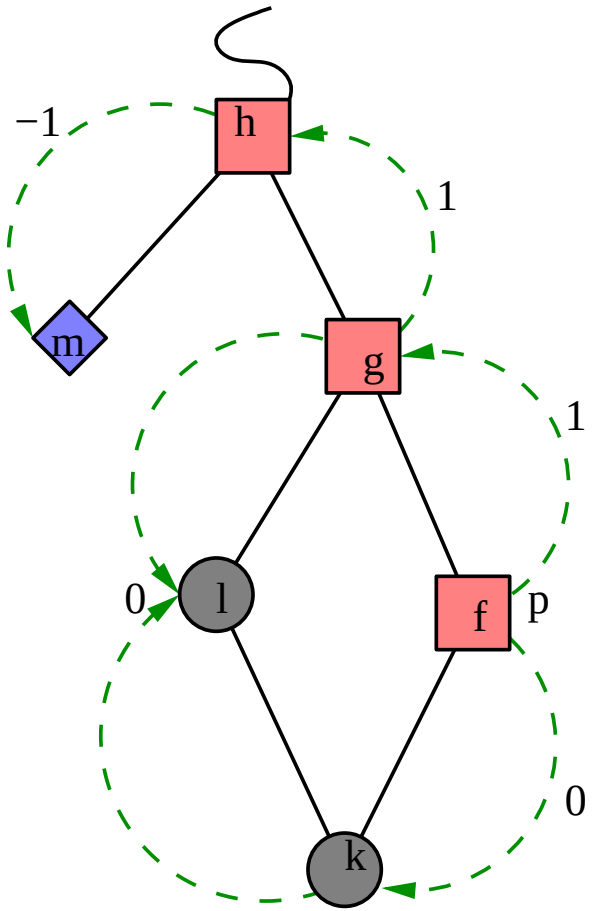
- Assigning node states correctly is not a trivial manner.
- We must respect/exploit GO's hierarchical structure .



- What is state of gene p with respect to function
  - f: 1
  - g: 1
  - h: 1
  - m: -1
  - k: 0
  - l: -1 or 0?

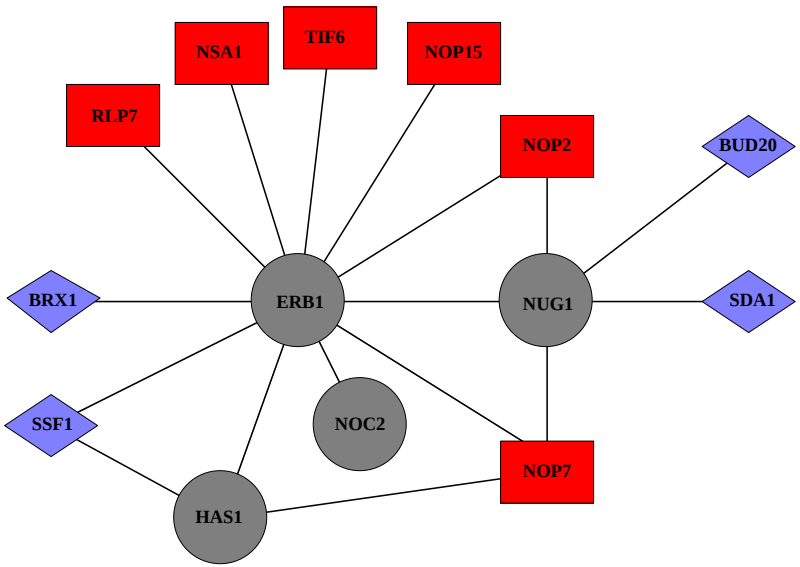
Assigning Node States

- Assigning node states correctly is not a trivial manner.
- We must respect/exploit GO's hierarchical structure .



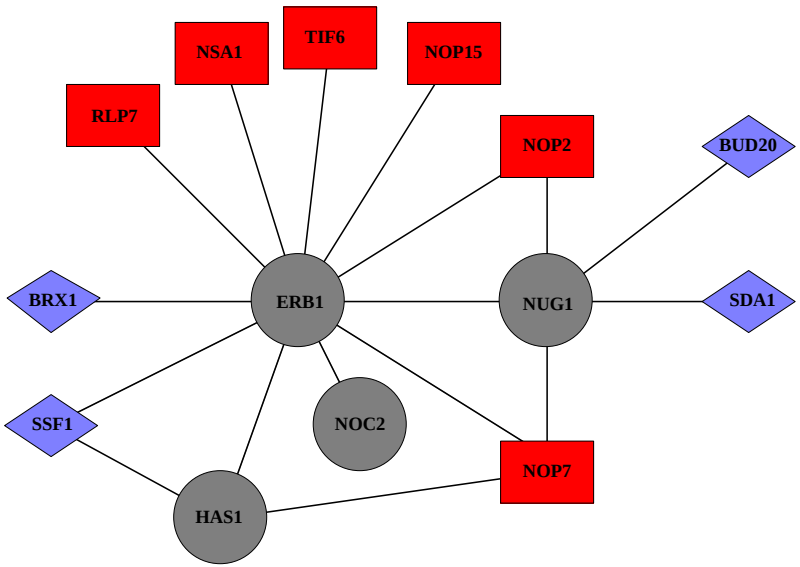
- What is state of gene p with respect to function
  - f: 1
  - g: 1
  - h: 1
  - m: -1
  - k: 0
  - l: -1 or 0? Correct state is 0.

Goal: Maximally-Consistent Assignments



- An edge is *consistent* if it is incident on nodes with the same state.
- *Maximally-consistent assignment*: number of consistent edges is maximised.

Goal: Maximally-Consistent Assignments



- An edge is *consistent* if it is incident on nodes with the same state.
- *Maximally-consistent assignment*: number of consistent edges is maximised.

Computational goal: Assign state of -1 or $+1$ to nodes with initial state 0 to achieve maximal consistency by minimising

$$E = -\frac{1}{2}\sum_i \sum_j w_{ij} s_i s_j$$

Predict nodes in state 1 as being annotated with the function.

Minimising E

- Finding state assignments to all nodes with initial $s_i = 0$ to minimise E is NP-complete if some edge weights are negative.

Minimising E

- Finding state assignments to all nodes with initial $s_i = 0$ to minimise E is NP-complete if some edge weights are negative.
- Vasquez et al., *Nature Biotech. 2003* use a simulated annealing-based approach.

Minimising E

- Finding state assignments to all nodes with initial $s_i = 0$ to minimise E is NP-complete if some edge weights are negative.
- Vasquez et al., *Nature Biotech. 2003* use a simulated annealing-based approach.
- Our approach is based on the idea of *local updates*: each node looks at its neighbours and decides what its state should be.

Minimising E

- Finding state assignments to all nodes with initial $s_i = 0$ to minimise E is NP-complete if some edge weights are negative.
- Vasquez et al., *Nature Biotech. 2003* use a simulated annealing-based approach.
- Our approach is based on the idea of *local updates*: each node looks at its neighbours and decides what its state should be.
- Both approaches are well-known and well-studied.

Minimising E

- Finding state assignments to all nodes with initial $s_i = 0$ to minimise E is NP-complete if some edge weights are negative.
- Vasquez et al., *Nature Biotech. 2003* use a simulated annealing-based approach.
- Our approach is based on the idea of *local updates*: each node looks at its neighbours and decides what its state should be.
- Both approaches are well-known and well-studied.
- Can use minimum cuts and integer programming (Nabieva et al., Proc. ISMB 2005; Murali, Wu, and Kasif, *Nature Biotech.*, 2006).

Local Update Rule

- Activation rule is

$$s_i = \mathrm{sgn}\left(\sum_{j \in N_i} w_{ij} s_j \right),$$

where N_i = neighbours of node i.

Local Update Rule

- Activation rule is

$$s_i = \mathrm{sgn}\left(\sum_{j \in N_i} w_{ij} s_j \right),$$

  where N_i = neighbours of node i.
- Applying this rule:

Local Update Rule

- Activation rule is

$$s_i = \mathrm{sgn}\left(\sum_{j \in N_i} w_{ij} s_j \right),$$

  where N_i = neighbours of node i.
- Applying this rule:
  - Parallel update: each node updates itself in parallel with the other nodes.

Local Update Rule

- Activation rule is

$$s_i = \mathrm{sgn}\left(\sum_{j \in N_i} w_{ij} s_j \right),$$

  where N_i = neighbours of node i.
- Applying this rule:
  - Parallel update: each node updates itself in parallel with the other nodes.
  - Serial update: go through each node in sequence.

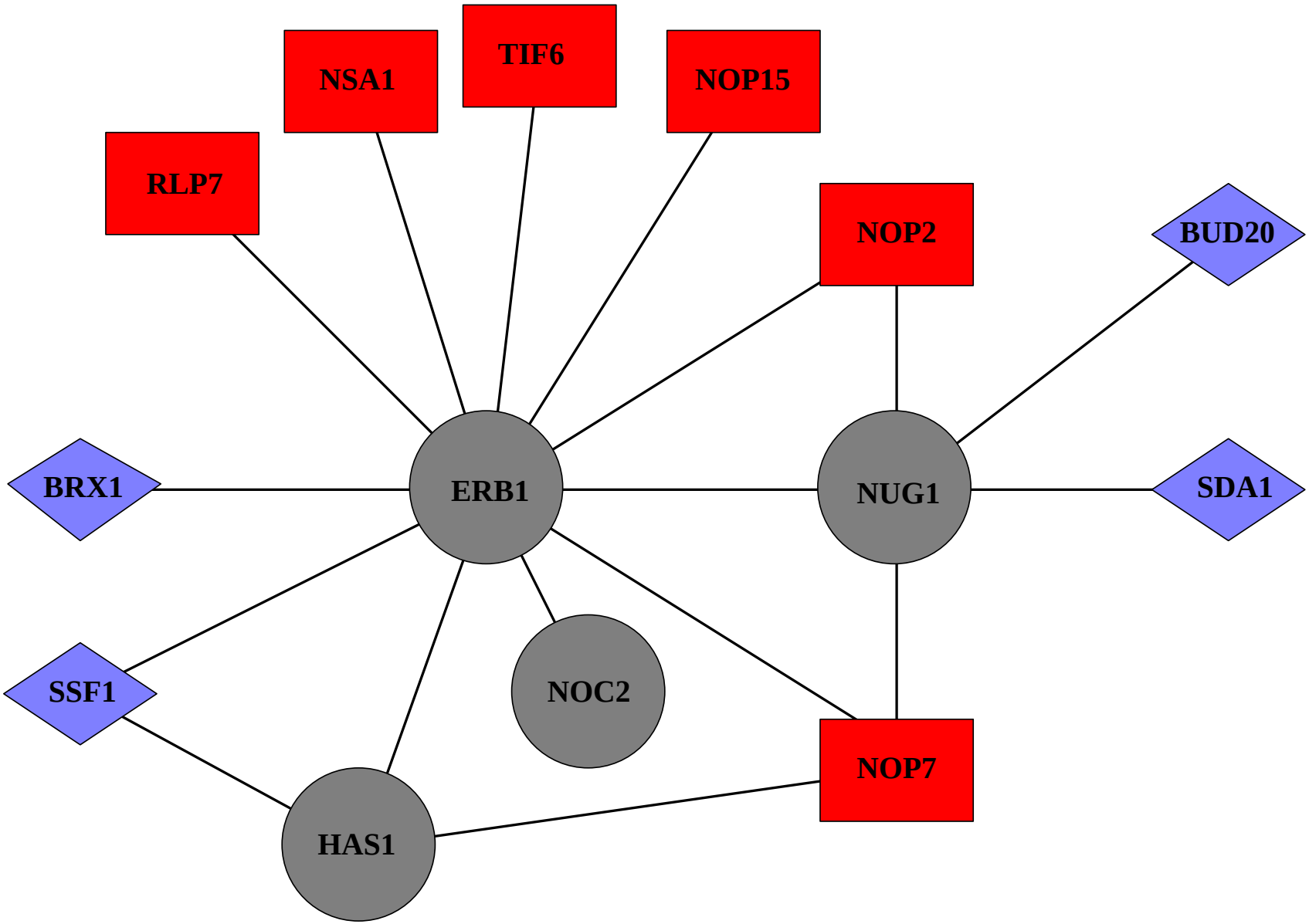
Local Update Rule

- Activation rule is

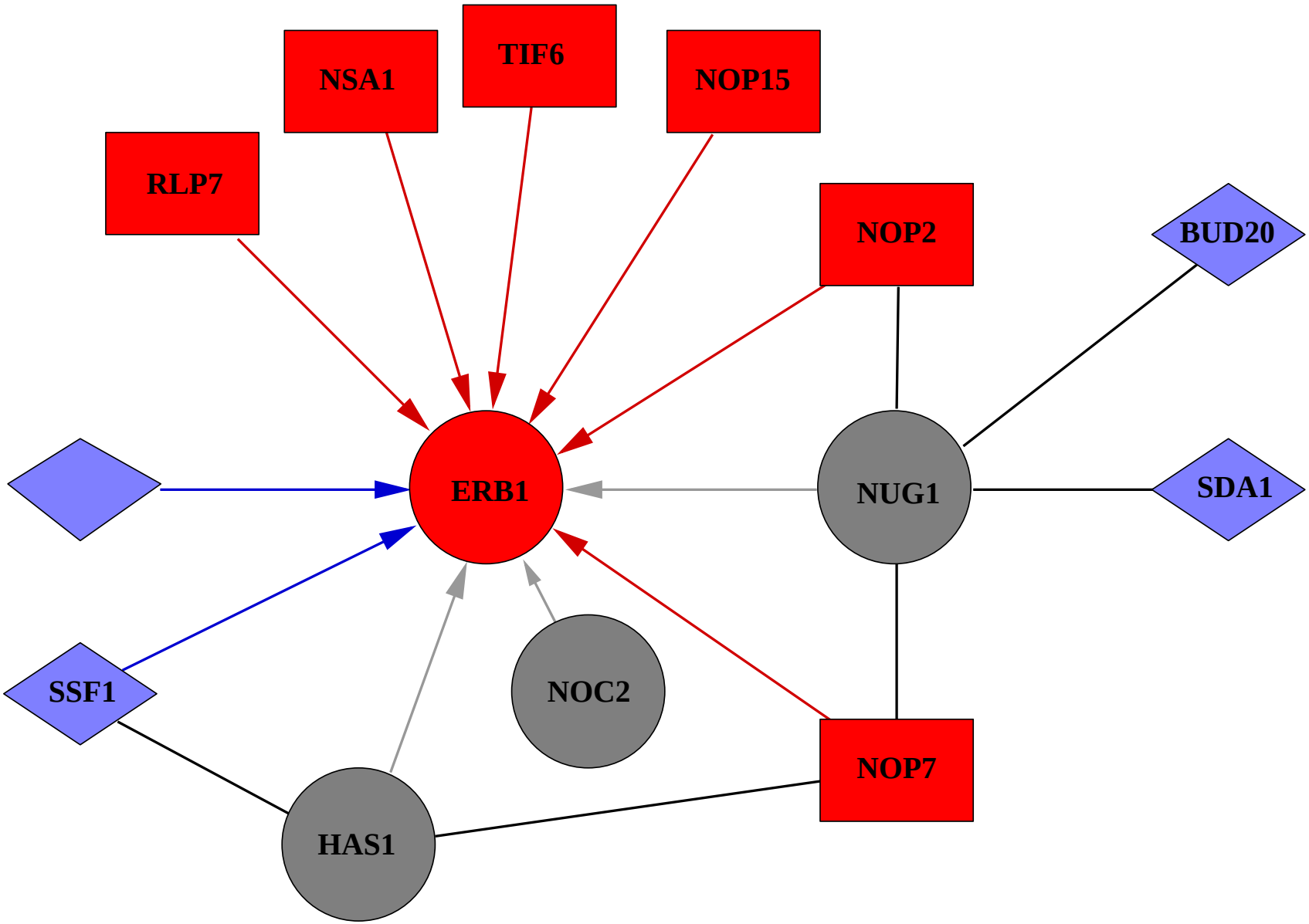
$$s_i = \mathrm{sgn}\left(\sum_{j \in N_i} w_{ij} s_j \right),$$

  where N_i = neighbours of node i.
- Applying this rule:
  - Parallel update: each node updates itself in parallel with the other nodes.
  - Serial update: go through each node in sequence.
- Stopping criterion: converge when no node's state changes.

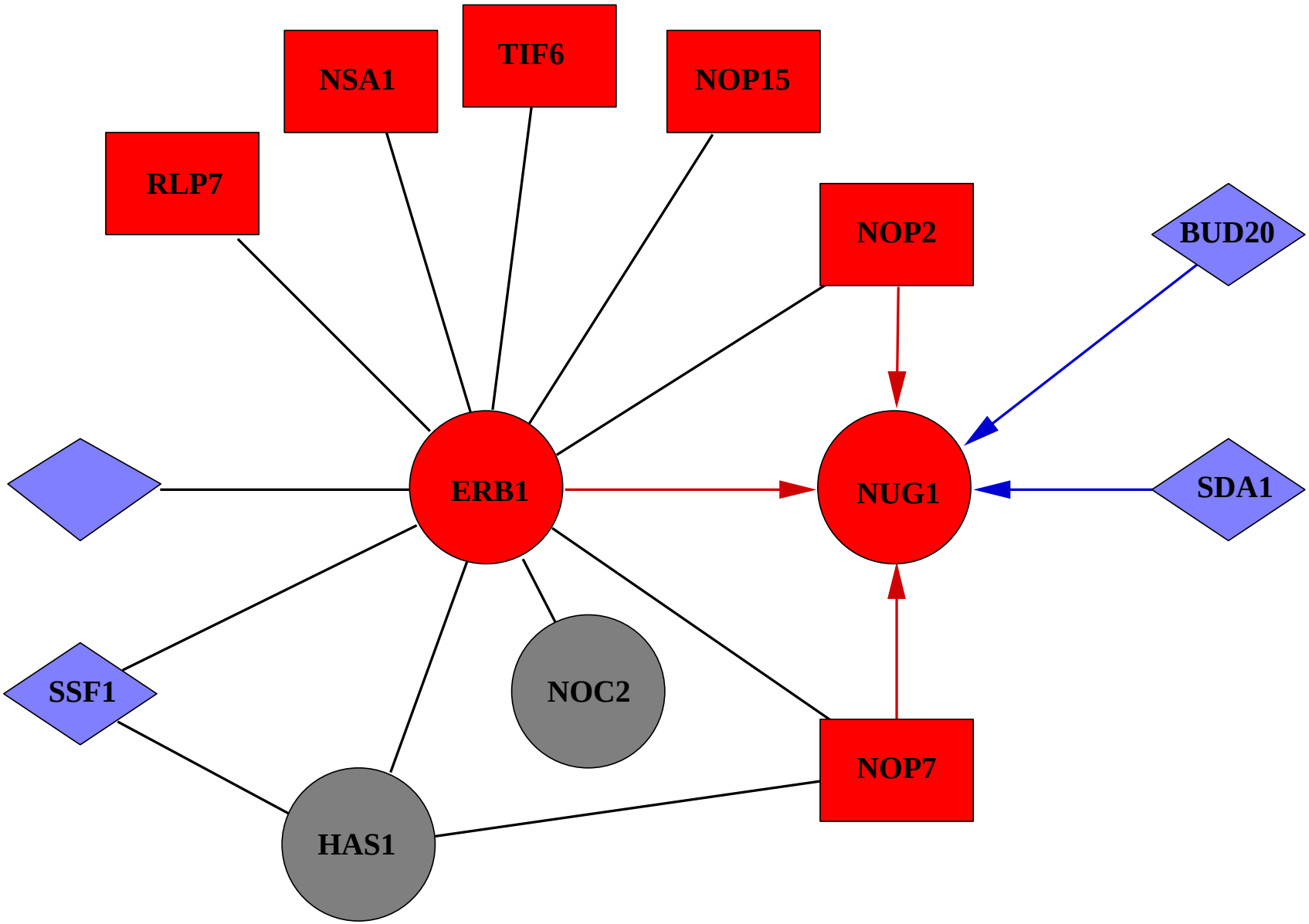
Example of Local Updates



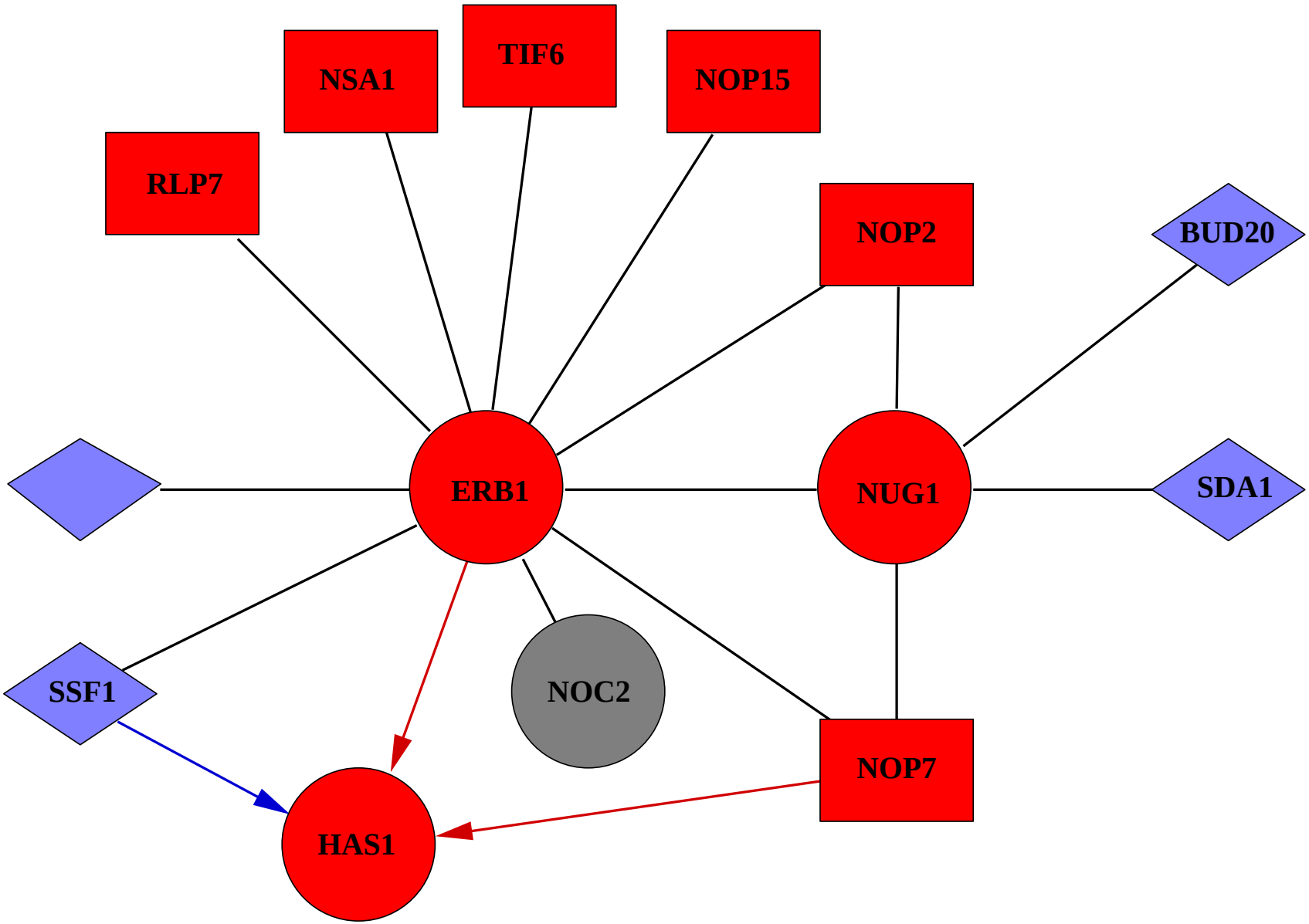
Example of Local Updates



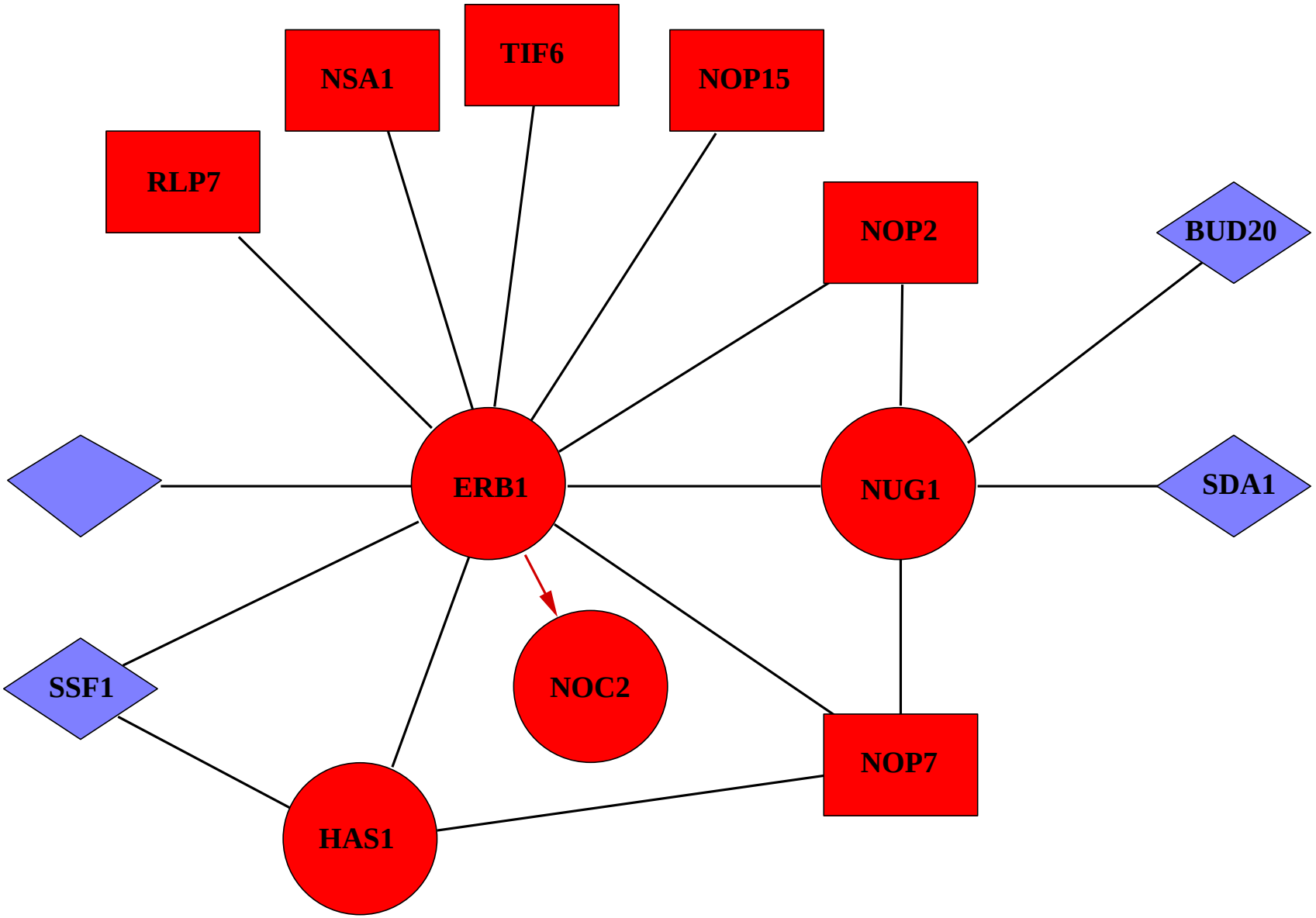
Example of Local Updates



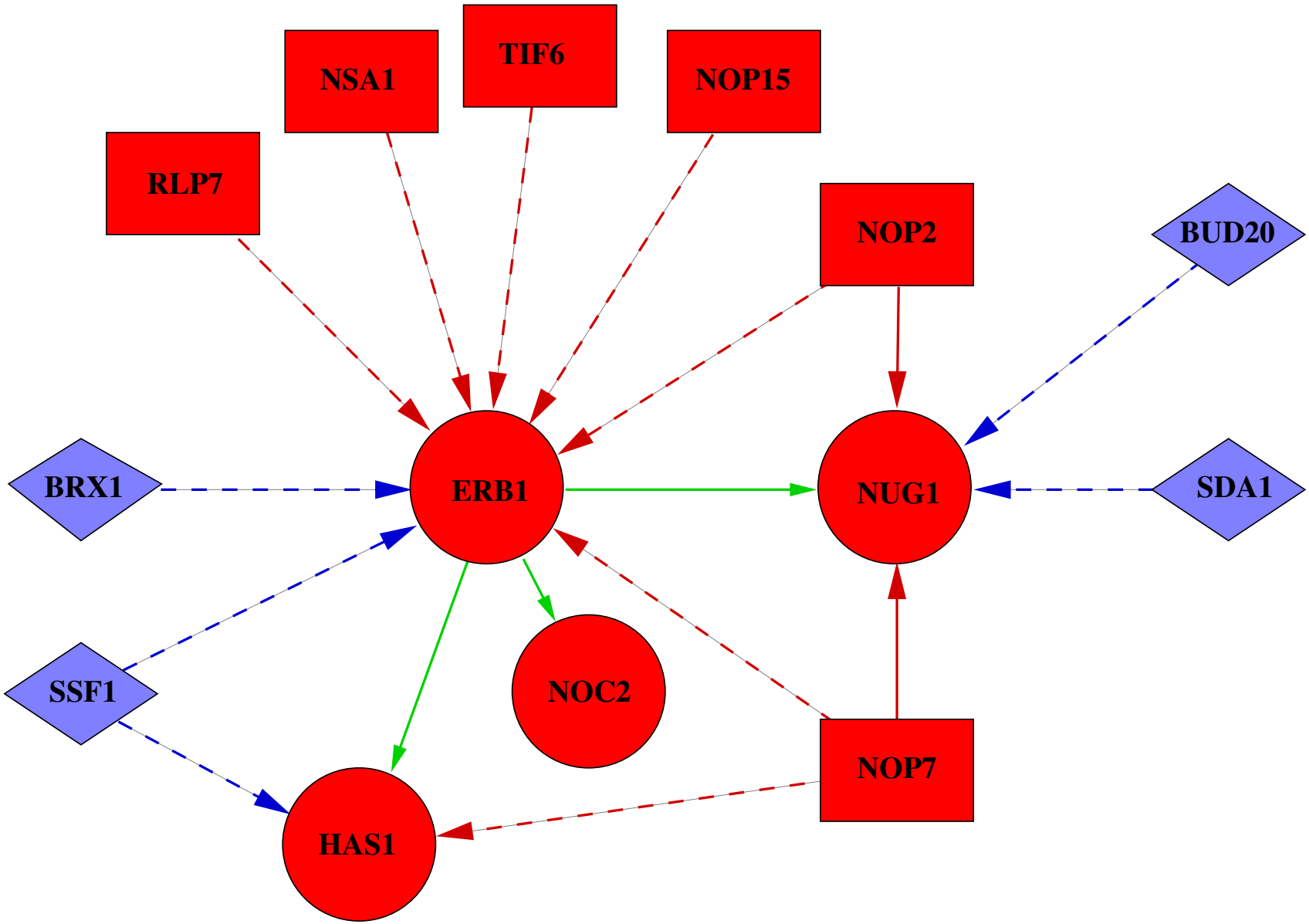
Example of Local Updates



Example of Local Updates



Example of Local Updates



Data Sets

- Interactions: General Repository of Interaction Datasets (GRID).
- Microarray: *Functional discovery via a compendium of expression profiles*. Hughes TR et al. *Cell*. 2000 102: 109–26.
- Functional Annotations: Gene Ontology, three categories are biological process, molecular function, and cellular component.

Cleaning Up PPI Network

- GRID data set has 4711 genes and 13607 interactions.
- GRID data set has information on publications.

ORF_A	ORF_B	EXPERIMENTAL_SYSTEM	SOURCE	PUBMED_ID
YER006W	YPL211W	Affinity Precipitation	Bassler et al.	;11583615;
YDL140C	YBR154C	Two Hybrid	BIND	;2496296;9207794;10393904;

- We only consider interactions reported by at least two different experiments to obtain 997 interactions between 1004 genes.

Data Integration

- Unweighted: $w_{ij} = 1$.
- Integrated: w_{ij} is the absolute value of correlation coefficient of the expression profiles of gene i and gene j in the "Compendium" data set.

Evaluation

- Leave one-out cross validation: For each function f,
  1. for each gene i annotated with f, set initial value of $s_i = 0$ and compute state assigned to i by the Hopfield network.
  2. Perform a similar operation for each gene not annotated with f.

Evaluation

- Leave one-out cross validation: For each function f,
  1. for each gene i annotated with f, set initial value of $s_i = 0$ and compute state assigned to i by the Hopfield network.
  2. Perform a similar operation for each gene not annotated with f.
- Measurement:

- True positive: $s_i : 1 \rightarrow 0 \rightarrow 1$
- False positive: $s_i : -1 \rightarrow 0 \rightarrow 1$
- True negative: $s_i : -1 \rightarrow 0 \rightarrow -1$
- False negative: $s_i : 1 \rightarrow 0 \rightarrow -1$

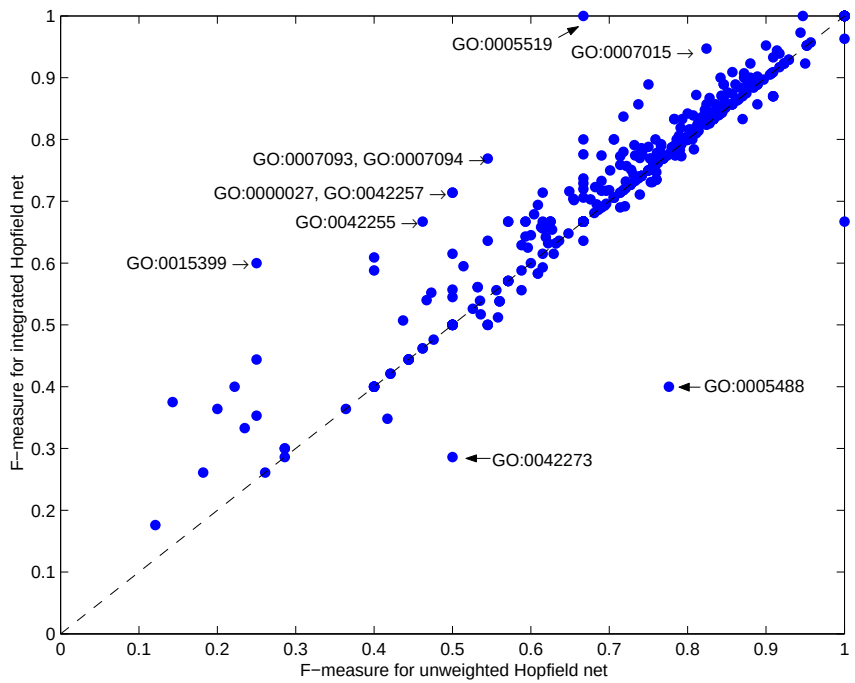
- Precision $=$ TP/(TP $+$ FP)
- Sensitivity $=$ Recall $=$ TP/(TP $+$ FN)
- F-measure $=$ Harmonic mean of precision and recall.

Results for Both Variants

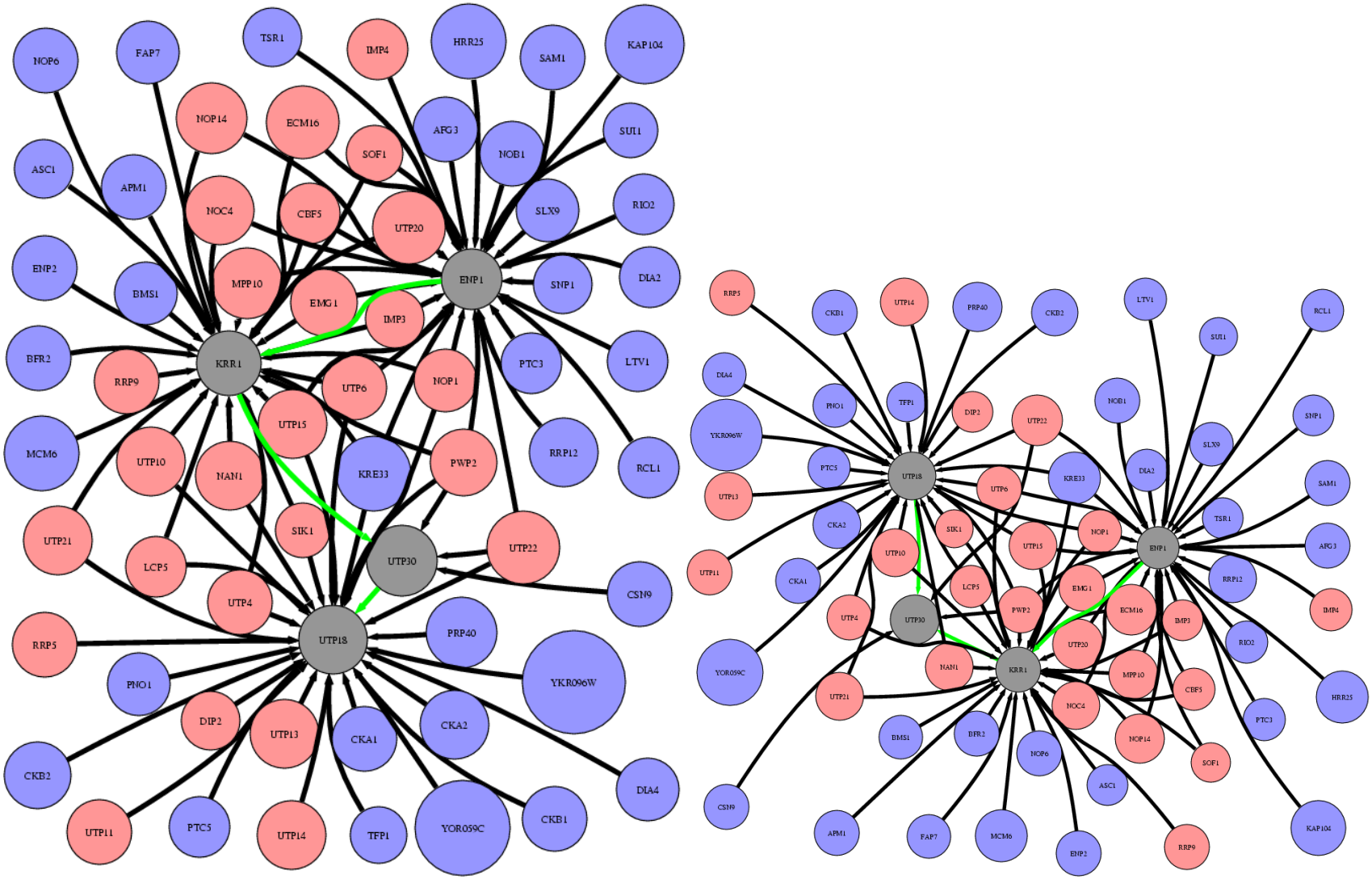
1. Overall comparison of cross-validation.
2. Specific examples of genes that perform better on cross-validation (see paper).
3. Novel functional annotations.
4. Propagation diagrams.

Overall Cross-Validation Results

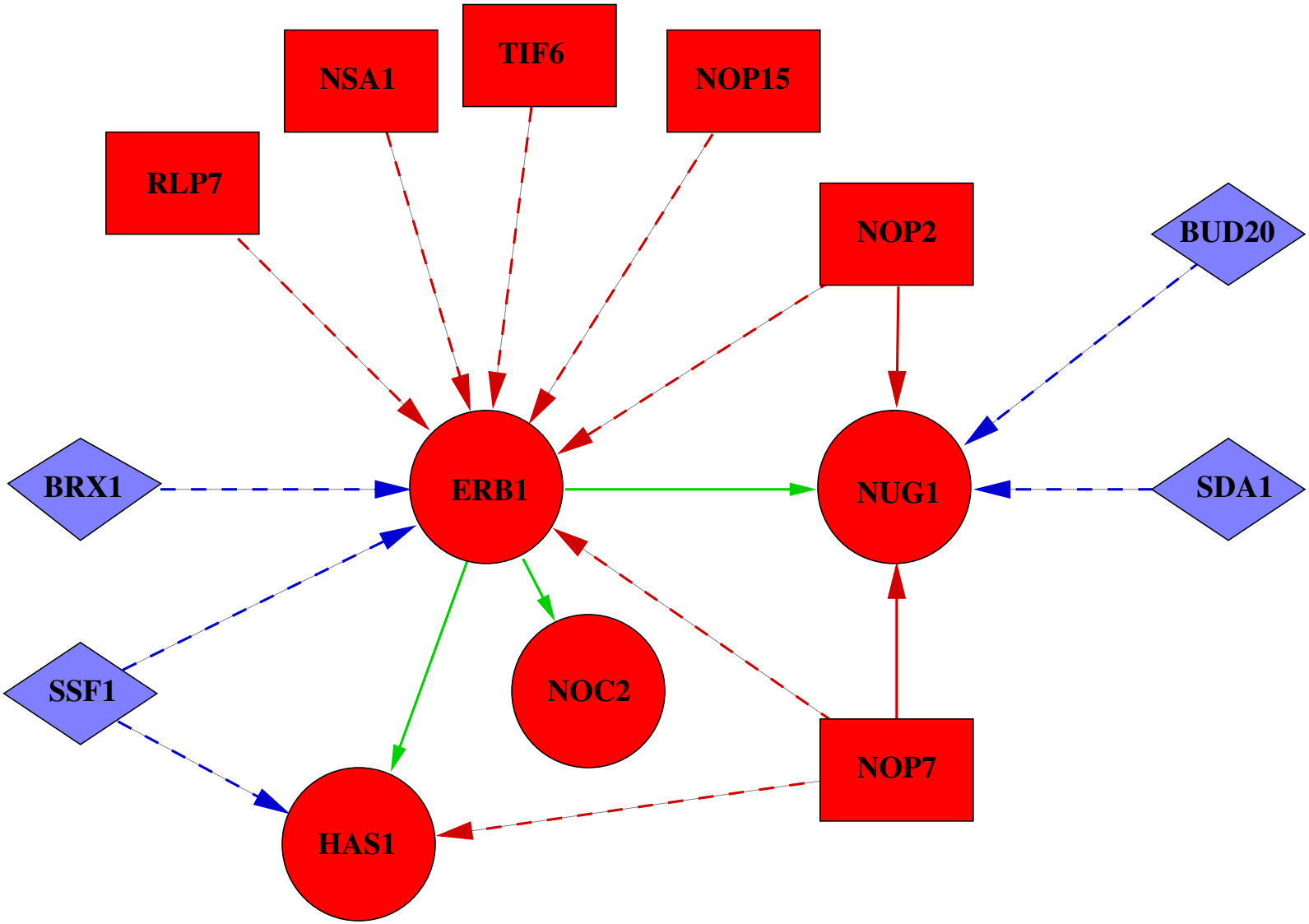
- Restricted to 828 functions for which F-score > 0.
- Unweighted network: Precision $= 94\%$, Recall $= 64\%$.
- Integrated network: Among 440 functions for which we make at least one novel prediction,
  - 168 function had better F-measures, 227 the same, and 45 smaller F-measures in the integrated network.



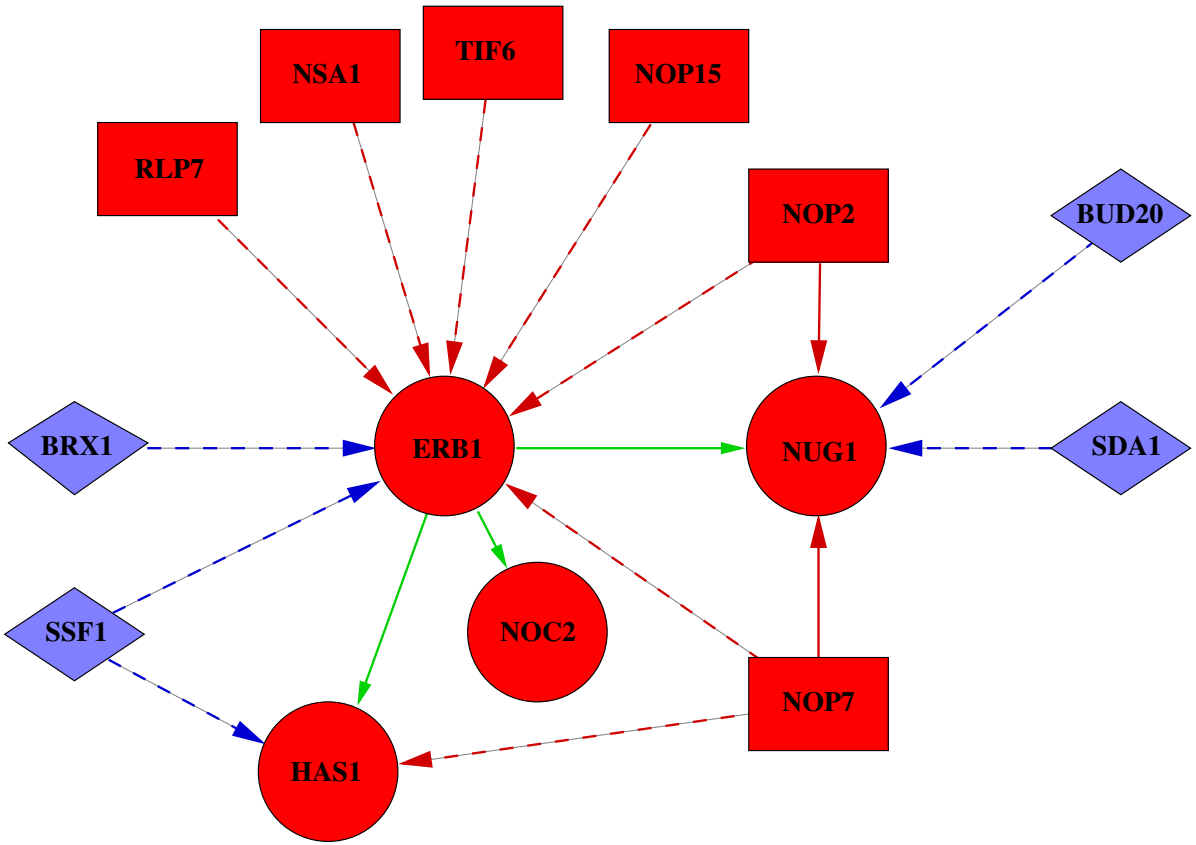
Propagation Diagrams



Propagation Diagrams



Novel Functional Annotations



- ERB1, HAS1, and NUG1: validated to have the function "rRNA processing."
- NOC2: validated to have the function "ribosome assembly and ribosome-nuclear export."

Novel Functional Annotations

- NHP10
  - biological process *chromatin modeling* and cellular component *chromatin remodeling complex*.
  - HMG1 proteins are involved in chromatin structure.
- UFO1
  - cellular component *nuclear ubiquitin ligase complex*
  - molecular function *ubiquitin-protein ligase activity* and biological processes *ubiquitin-dependent protein catabolism*.
- PKC1
  - cellular component *1,3 beta-glucan synthase complex*.
  - known: cellular component *intracellular* and biological processes *cell wall organization and biogenesis*.

More Novel Functional Annotations

- YKL067W
  - biological process *signal transduction* and cellular component *spindle pole body*.
  - molecular function *nucleoside-diphosphate kinase (NDK) activity*; NDK interferes with the mating pheromone signal transduction in *S. pombe*.
- YCR099C and YBL059W
  - biological process *ER to Golgi transport* and cellular component *COPII vesicle coat*.
  - Vesicles with COPII coats are found associated with ER membranes at steady state.

Overall Correctness of Predictions

- 207 predictions for functions with F-score $> 75\%$.
- 15 predictions are correct.
- 11 predictions at distance 1 from true function.
- 49 predictions at distance 2 from true function.
- Remaining predictions not validated.
- Validated functions include nucleolus, chromatin remodeling complex, snoRNA binding, RNA binding, vesicle-mediated transport.

Features of the GAIN System

- Systematic algorithm for propagating evidence in an FLN.
- Clean separation between construction of functional links and prediction of function.
- For each function, predictions are maximally consistent.
- Each prediction associated with measures of confidence.
- Propagation diagrams provide intuitive visualisation of evidence flow.
- VIRGO webserver for invoking GAIN and querying and browsing its predictions.