EMBL
Heidelberg
Germany

STRING

Predicting protein networks from genomic context and external data

Lars Juhl Jensen (jensen@embl.de)

Christian von Mering (mering@embl.de)

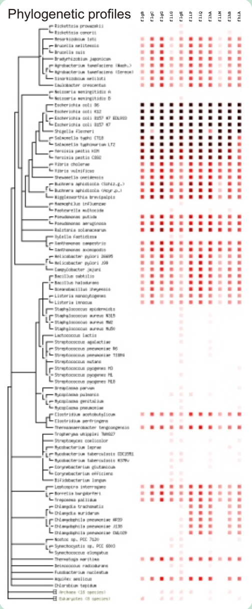
Peer Bork (bork@embl.de)

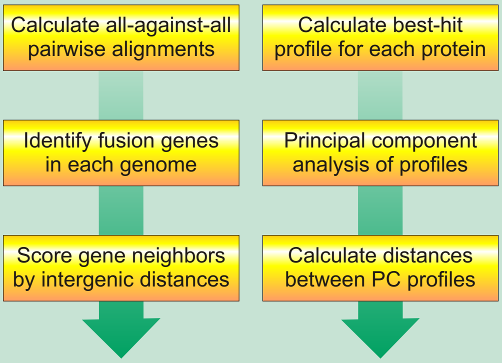
The web-based database STRING (http://string.embl.de) is a large, pre-computed resource that allows any protein of interest to be placed into a high-confidence network of functionally associated protein partners. For 100+ genomes, the tool evaluates genomic context, protein interaction data, gene co-expression, co-mentioning of genes in literature, and certain database annotations.

Genomic context

So-called genomic context methods allow discovery of functional relations from complete genome sequences. Two methods rely on functionally related genes often being arranged in operons (the gene neighborhood method) or even fused to a single gene (the gene fusion method). Finally, the genes of a functional module are typically either all present or all absent in each genome.

Genomic context methods tend to be most powerful for analyzing prokaryotic genomes. Part of the reason is that more prokaryotes than eukaryotes have been sequenced. The neighbourhood method, however, will never be powerful for eukaryotes, as it relies on detecting operons. In the near future, we intend to expand STRING with novel genomic context methods.




Calculate all-against-all pairwise alignments

Identify fusion genes in each genome

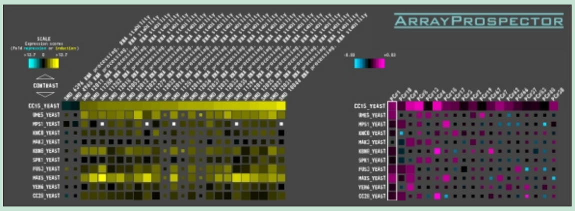
Score gene neighbors by intergenic distances

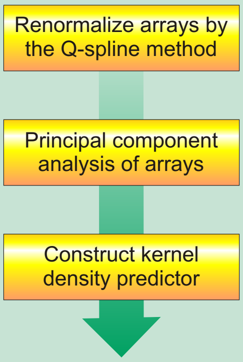
Calculate best-hit profile for each protein

Principal component analysis of profiles

Calculate distances between PC profiles

Large-scale experimental data




Renormalize arrays by the Q-spline method

Principal component analysis of arrays

Construct kernel density predictor

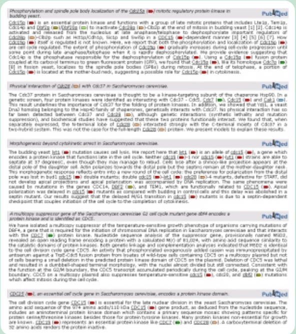
ArrayProspector is an interactive web resource of gene associations predicted from all expression data in the Stanford Microarray Database (SMD). The underlying pre-computed database currently contains more than 200,000 high confidence gene associations in 12 different species. The resource allows every association to be visually inspected. STRING utilizes both the evidence scores and the visualization capabilities of ArrayProspector.

In addition to microarray expression data, STRING also incorporates other types of large-scale experimental data sets, e.g. from yeast two-hybrid screens, complex purifications, and chromatin immunoprecipitation experiments. For all such data, a quality score is associated with every individual binary interaction.

http://www.bork.embl.de/ArrayProspector

Literature mining

Vast amounts of biological knowledge is buried in the scientific literature. STRING extracts protein associations from PubMed abstracts based on co-mentioning. We will soon also include more specific relations extracted by shallow parsing.


Phosphorylation and spindle pole body localization of the Cdc15p (■) mitotic regulatory protein kinase in budding yeast. PubMed

Cdc15p (■) is an essential protein kinase and functions with a group of late mitotic proteins that includes Lte1p, Tem1p, Cdc14p and Dbf2p (■)/Dbf20p (■) to inactivate Cdc28p (■)-Clb2p at the end of mitosis in budding yeast [1] [2]. Cdc14p is activated and released from the nucleolus at late anaphase/telophase to dephosphorylate important regulators of Cdc28p (■)-Clb2p such as Hct1p/Cdh1p, Sic1p and Swi5p in a CDC15 (■)-dependent manner [3] [4] [5] [6] [7]. How Cdc15p (■) itself is regulated is not known. Here, we report that both the phosphorylation and localization of Cdc15p (■) are cell cycle regulated. The extent of phosphorylation of Cdc15p (■) gradually increases during cell-cycle progression until some point during late anaphase/telophase when it is rapidly dephosphorylated. We provide evidence suggesting that Cdc14p is the phosphatase responsible for the dephosphorylation of Cdc15p (■). Using a Cdc15p (■) fusion protein coupled at its carboxyl terminus to green fluorescent protein (GFP), we found that Cdc15p (■), like its homologue Cdc7p (■) [8] in fission yeast, localizes to the spindle pole bodies (SPBs) during mitosis. At the end of telophase, a portion of Cdc15p (■) is located at the mother-bud neck, suggesting a possible role for Cdc15p (■) in cytokinesis.

Physical interaction of Cdc28 (■) with Cdc37 in Saccharomyces cerevisiae. PubMed

The Cdc37 protein in Saccharomyces cerevisiae is thought to be a kinase-targeting subunit of the chaperone Hsp90. In a genetic screen, four protein kinases were identified as interacting with Cdc37 - Cdc5, Cdc7 (■), Cdc15 (■) and Cak1 (■). This result underlines the importance of Cdc37 for the folding of protein kinases. In addition, we showed that Ydj1, a yeast DnaJ homolog belonging to the Hsp40 family of chaperones, genetically interacts with Cdc37. No physical interaction has so far been detected between Cdc37 and Cdc28 (■), although genetic interactions (synthetic lethality and mutation suppression), and biochemical studies have suggested that these two proteins functionally interact. We found that, when separately expressed, the N-terminal lobe of Cdc28 (■) interacted strongly with the C-terminal moiety of Cdc37 in a two-hybrid system. This was not the case for the full-length Cdc28 (■) protein. We present models to explain these results.

Morphogenesis beyond cytokinetic arrest in Saccharomyces cerevisiae. PubMed

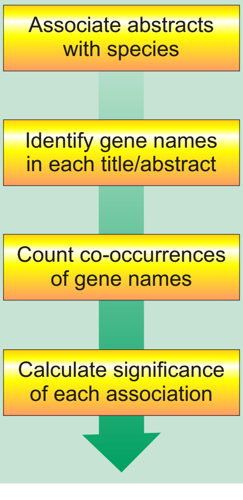
The budding yeast lyt1 (■) mutation causes cell lysis. We report here that lyt1 (■) is an allele of cdc15 (■), a gene which encodes a protein kinase that functions late in the cell cycle. Neither cdc15 (■)-1 nor cdc15 (■)-lyt1 (■) strains are able to septate at 37 degreesC, even though they may manage to rebud. Cells lyse after a shmoo-like projection appears at the distal pole of the daughter cell. Actin polarizes towards the distal pole but the septins remain at the mother-daughter neck. This morphogenetic response reflects entry into a new round of the cell cycle: the preference for polarization from the distal pole was lost in bud1 cdc15 (■) double mutants; double cdc15 (■)-lyt1 (■) cdc28 (■)-4 mutants, defective for START, did not develop apical projections and apical polarization was accompanied by DNA replication. The same phenomena were caused by mutations in the genes CDC14, DBF2 (■), and TEM1, which are functionally related to CDC15 (■). Apical polarization was delayed in cdc15 (■) mutants as compared with budding in control cells and this delay was abolished in a septin mutant. Our results suggest that the delayed M/G1 transition in cdc15 (■) mutants is due to a septin-dependent checkpoint that couples initiation of the cell cycle to the completion of cytokinesis.

A multicopy suppressor gene of the Saccharomyces cerevisiae G1 cell cycle mutant gene dbf4 encodes a protein kinase and is identified as CDC5. PubMed

We have isolated a multicopy suppressor of the temperature-sensitive growth phenotype of organisms carrying mutations of DBF4, a gene that is required for the initiation of chromosomal DNA replication in Saccharomyces cerevisiae and that interacts with the CDC7 (■) protein kinase. Nucleotide sequence analysis of the suppressor gene, provisionally named MSD2, revealed an open reading frame encoding a protein with a calculated M(r) of 81,024, with amino acid sequence similarity to the catalytic domains of protein kinases. Both genetic linkage and complementation analyses indicated that MSD2 is identical to the cell division cycle gene CDC5. An activity that phosphorylated exogenously added casein was immunoprecipitated by antiserum against a TrpE-Cdc5 fusion protein from lysates of wild-type cells containing CDC5 on a multicopy plasmid but not of cells bearing a small deletion in the predicted protein kinase domain of CDC5 on the plasmid. Deletion of CDC5 was lethal and resulted in a dumbbell-shaped terminal morphology, with the nuclei almost divided but still connected. Consistent with the function at the G2/M boundary, the CDC5 transcript accumulated periodically during the cell cycle, peaking at the G2/M boundary. CDC5 on a multicopy plasmid also suppresses temperature-sensitive cdc15 (■), cdc20, and dbf2 (■) mutations which affect mitosis during the cell cycle.

CDC15 (■), an essential cell cycle gene in Saccharomyces cerevisiae, encodes a protein kinase domain. PubMed

The cell division cycle gene CDC15 (■) is essential for the late nuclear division in the yeast Saccharomyces cerevisiae. The amino acid sequence of the 974 amino acids/110 kDa CDC15 (■) gene product, as deduced from the nucleotide sequence, includes an aminoterminal protein kinase domain which contains a primary sequence mosaic showing patterns specific for protein serine/threonine kinases besides those for protein tyrosine kinases. Many protein kinases non-essential for growth are known. CDC15 (■) represents an essential protein kinase like CDC7 (■) and CDC28 (■). A carboxyterminal deletion of 32 amino acids renders the protein inactive.

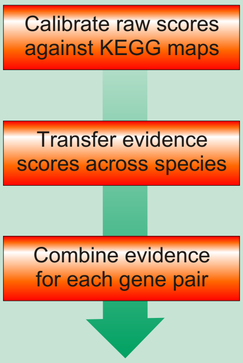

Associate abstracts with species

Identify gene names in each title/abstract

Count co-occurrences of gene names

Calculate significance of each association

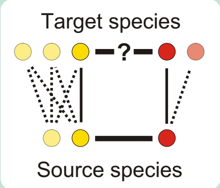
Integration of diverse evidence


Calibrate raw scores against KEGG maps

Transfer evidence scores across species

Combine evidence for each gene pair

The notion of "functional relations" is central to STRING. All raw evidence scores are calibrated against KEGG metabolic maps, which serves as our reference set. Calibrated scores for different evidence types are directly comparable and easily combined.



All types of evidence in STRING are transferred across species using a "fuzzy orthology" scheme.

Related publications

C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel
"STRING: A database of predicted functional associations between proteins"
Nucleic Acids Research, **31**, 258-261, 2003

J.O. Korbel, L.J. Jensen, C. von Mering and P. Bork
"Analysis of genomic context: Prediction of functional associations from bidirectionally transcribed gene pairs"
Nature Biotechnology, **22**, 911-917, 2004

L.J. Jensen, J. Lagarde, C. von Mering and P. Bork
"ArrayProspector: A web resource of functional associations inferred from microarray expression data"
Nucleic Acids Research, **32**, W445-W448, 2004

L.J. Jensen and P. Bork
"Quality analysis and integration of large-scale molecular data sets"
Drug Discovery Today: TARGETS, **3**, 51-56, 2004

P. Bork, L.J. Jensen, C. von Mering, A.K. Ramani, I. Lee and E.M. Marcotte
"Protein interaction networks from yeast to human"
Current Opinions in Structural Biology, **14**, 292-299, 2004

J. Saric, L.J. Jensen, R. Ouzounova, I. Rojas and P. Bork
"Extracting regulatory gene expression networks from PubMed"
in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004

A network of functional modules

Through the cross-species integration of many, diverse evidence types, STRING provides a network of functional association for proteins from 100+ species. This entire network can be browsed online at http://string.embl.de

The network is highly ordered, and functional modules can be discovered by performing e.g. means clustering of the network. The resulting functional modules are not isolated, but rather form a network of their own.

To the right is shown a small example network of three related functional modules from yeast: the ATP synthetase complex, the cytochrome C oxidase complex and the ubiquitinol-cytochrome C reductase complex.