Algorithm Engineering for Color-Coding to Facilitate Signaling Pathway Detection

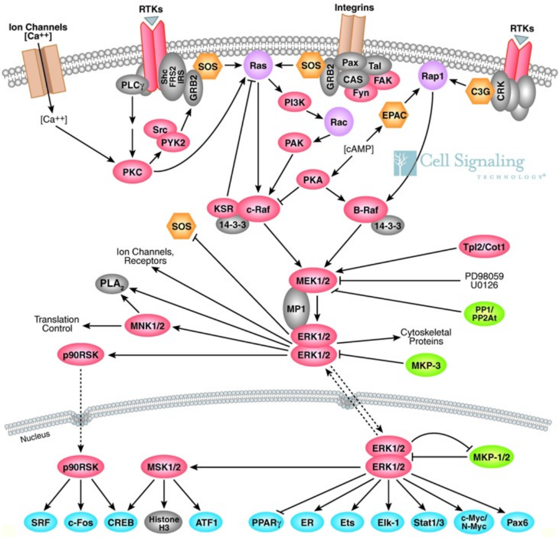
Falk Hüffner Sebastian Wernicke Thomas Zichner

Friedrich-Schiller-Universität Jena

Fifth Asia Pacific Bioinformatics Conference
January 17, 2007

Outline

Protein Interaction Networks



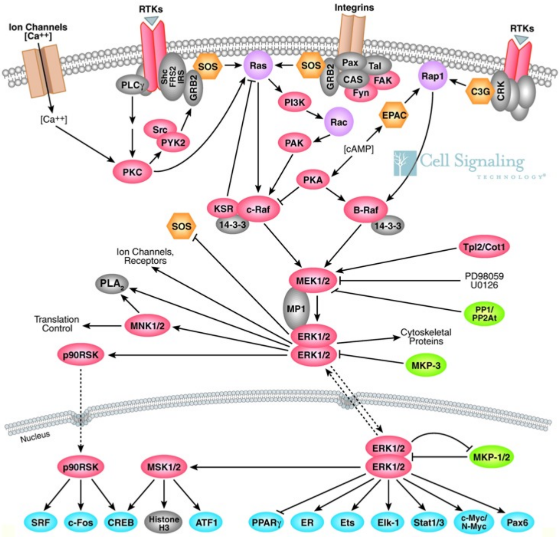
[www.cellsignal.com]

Protein Interaction Networks

Representation of protein interactions as a graph:

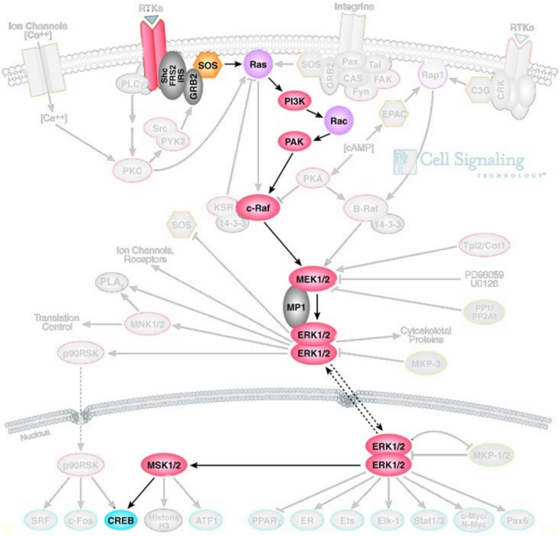
- Proteins are nodes
- Interactions are edges
- Edges are annotated with interaction probability (obtained by two-hybrid screening)

Signaling Pathways



[www.cellsignal.com]

Signaling Pathways



[www.cellsignal.com]

Signaling Pathways

Sequence of distinct proteins, where each interacts strongly with the previous one.

Most Probable Path

Input: Graph $G = (V, E)$, interaction probabilities $p : E \longrightarrow [0, 1]$, integer $k > 0$.
Task: Find a non-overlapping path $v_1, \ldots, v_k$ of length k in G that maximizes $p(v_1, v_2) \cdot \ldots \cdot p(v_{k-1}, v_k)$.

Signaling Pathways

Sequence of distinct proteins, where each interacts strongly with the previous one.

Most Probable Path

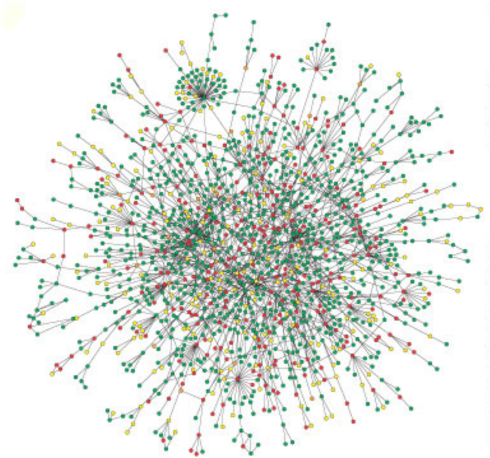
Input: Graph $G = (V, E)$, interaction probabilities $p : E \to [0, 1]$, integer $k > 0$.
Task: Find a non-overlapping path $v_1, \ldots, v_k$ of length k in G that maximizes $p(v_1, v_2) \cdot \ldots \cdot p(v_{k-1}, v_k)$.

Setting $w(e) := -\log(p(e))$:

Minimum-Weight Path

Input: Graph $G = (V, E)$, weights $w : E \to [0, 1]$, integer $k > 0$.
Task: Find a non-overlapping path $v_1, \ldots, v_k$ of length k in G that minimizes $w(v_1, v_2) + \cdots + w(v_{k-1}, v_k)$.

Yeast Network



4 400 proteins, 14 300 interactions, looking for paths of length 5–15

Minimum-Weight Path

Theorem

MINIMUM-WEIGHT PATH *is NP-hard [GAREY&JOHNSON 1979].*

For an exact algorithm, we have to accept exponential runtime.

Idea

Exploit the fact that the paths sought for are rather short ($\approx$ 5–15): restrict the exponential part of the runtime to k (parameterized complexity).

Color-Coding

Color-coding [Alon, Yuster&Zwick J. ACM 1995]:

- randomly color each vertex of the graph with one of k colors
- hope that all vertices in the subgraph searched for obtain different colors (colorful)
- solve the Minimum-Weight Path under this assumption (which is much quicker)
- repeat until it is reasonably certain that the path was colorful at least once

Result: exponential part of the runtime depends only on k

Dynamic Programming for Minimum-Weight Colorful Path

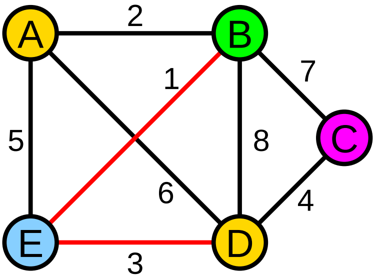
Idea

Table entry $W[v, C]$ stores the minimum-weight path that ends in v and uses exactly the colors in S.

Dynamic Programming for Minimum-Weight Colorful Path

Idea

Table entry $W[v, C]$ stores the minimum-weight path that ends in v and uses exactly the colors in S.



$W[B, \{\bigcirc, \bigcirc, \bigcirc\}] = 4$

Dynamic Programming for Minimum-Weight Colorful Path

Coloring $c : V \to \{1, \ldots, k\}$

Recurrence

$$W[v, C] = \min_{u \in N(v) | c(u) \in C \setminus \{c(v)\}} (W[u, C \setminus \{c(v)\}] + w(u, v))$$

Dynamic Programming for Minimum-Weight Colorful Path

Coloring $c : V \to \{1, \ldots, k\}$

Recurrence

$$W[v, C] = \min_{u \in N(v) | c(u) \in C \setminus \{c(v)\}} (W[u, C \setminus \{c(v)\}] + w(u, v))$$

- Each table entry can be calculated in $O(n)$ time
- $n2^k$ table entries

$\leadsto$ Runtime: $O(n \cdot n2^k) = n^2 \cdot 2^k$

Color-coding Runtime

- $O(n^2 \cdot 2^k)$ time per trial
- To obtain error probability ε, one needs $O(|\ln \varepsilon| \cdot e^k)$ trials

Theorem ([Alon et al. JACM 1995])

Minimum-Weight Path *can be solved in* $O(|\ln \varepsilon| \cdot 5.44^k |G|)$ *time).*

Color-coding Runtime

- $O(n^2 \cdot 2^k)$ time per trial
- To obtain error probability ε, one needs $O(|\ln \varepsilon| \cdot e^k)$ trials

Theorem ([Alon et al. JACM 1995])

MINIMUM-WEIGHT PATH *can be solved in* $O(|\ln \varepsilon| \cdot 5.44^k |G|)$ *time).*

Color-coding can find minimum-weight paths of length 10 in the yeast protein interaction networks within 3 hours ($n = 4\,400$, $k = 10$) [Scott et al., RECOMB'05]

Increasing the Number of Colors

Idea

Use $k + x$ colors instead of k colors.

Trial runtime:

$$O(2^k|G|) \to O(2^{k+x}|G|)$$

Increasing the Number of Colors

Idea

Use $k + x$ colors instead of k colors.

Trial runtime:

$$O(2^k|G|) \to O(2^{k+x}|G|)$$

Probability P_c for colorful path ($k = 8$, $\varepsilon = 0.001$):

x	0	1	2	3	4	5
P_c	0.0024	0.0084	0.0181	0.0310	0.0464	0.0636
trials	2871	816	378	220	146	106

Increasing the Number of Colors

Idea

Use $k + x$ colors instead of k colors.

Trial runtime:

$$O(2^k|G|) \to O(2^{k+x}|G|)$$

Probability P_c for colorful path ($k = 8$, $\varepsilon = 0.001$):

x	0	1	2	3	4	5
P_c	0.0024	0.0084	0.0181	0.0310	0.0464	0.0636
trials	2871	816	378	220	146	106

Theorem

MINIMUM-WEIGHT PATH *can be solved in* $O(|\ln \varepsilon| \cdot 4.32^k|G|)$ *time by choosing* $x = 0.3k$.

Increasing the Number of Colors

Idea

Use $k + x$ colors instead of k colors.

Trial runtime:

$$O(2^k|G|) \to O(2^{k+x}|G|)$$

Probability P_c for colorful path ($k = 8$, $\varepsilon = 0.001$):

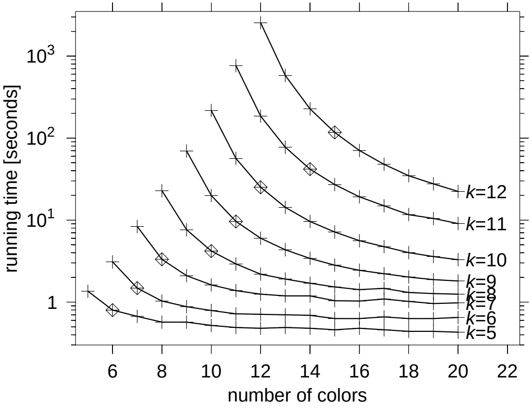
x	0	1	2	3	4	5
P_c	0.0024	0.0084	0.0181	0.0310	0.0464	0.0636
trials	2871	816	378	220	146	106

Theorem

MINIMUM-WEIGHT PATH *can be solved in* $O(|\ln\varepsilon| \cdot 4.32^k|G|)$ *time by choosing* $x = 0.3k$.

But: Higher memory usage

Increasing the Number of Colors



Runtimes for the yeast protein interaction network (highlighted point of each curve marks worst-case optimum)

Exploiting Lower Bounds

Idea

Use a known solution to prune “hopeless” table entries.

- Discard entries that already have a weight higher than the known solution.

Exploiting Lower Bounds

Idea

Use a known solution to prune "hopeless" table entries.

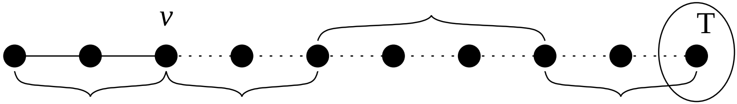
- Discard entries that already have a weight higher than the known solution.
- Discard entries when

$$\text{weight} + (\text{minimum edge weight} \cdot \text{edges left})$$

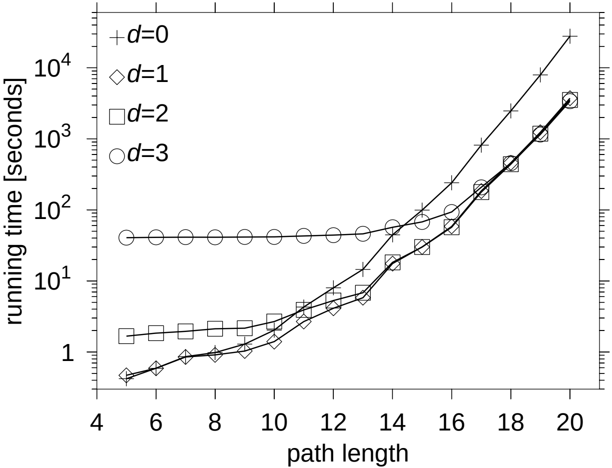
is higher than the weight of the known solution.

Precalculated Lower Bounds

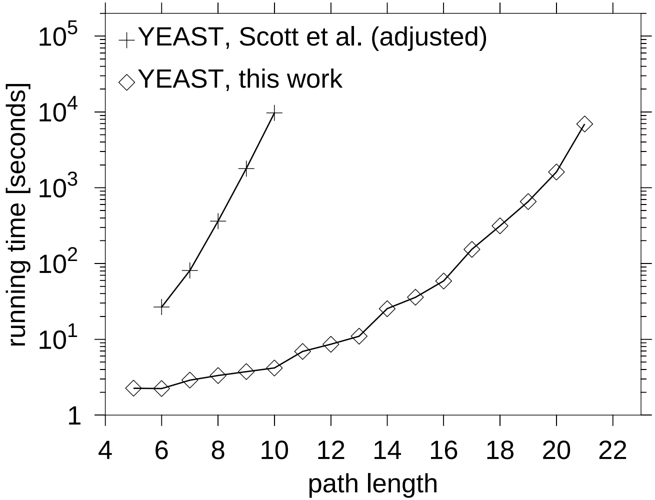
For each vertex u and a range of lengths $1 \leq i \leq d$, determine the minimum weight of a path of i edges that starts at u.



Lower Bounds Experiments



Yeast Network

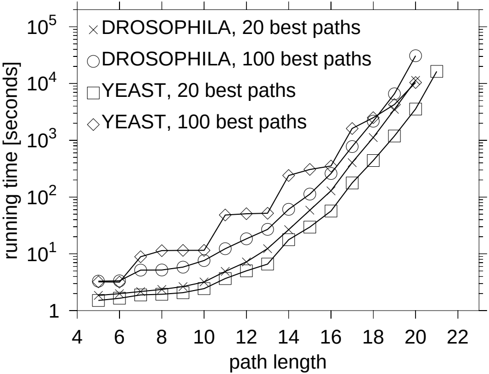


Network Comparison

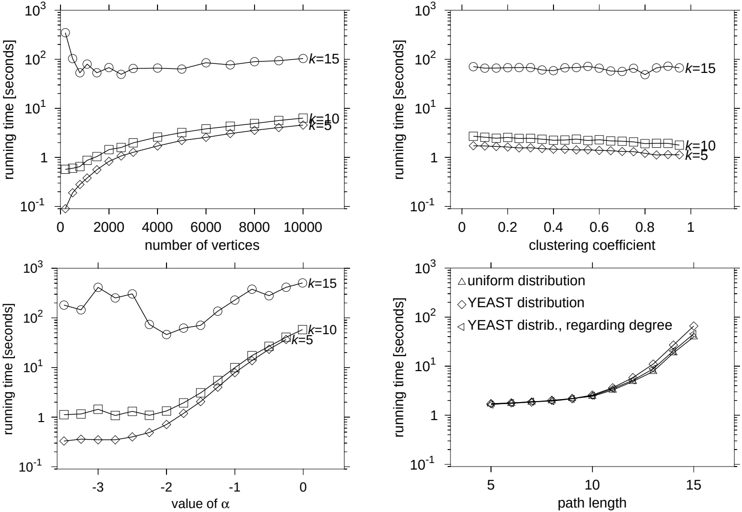
	$\lvert V \rvert$	$\lvert E \rvert$	clust. coeff.	avg. degree	max. degree
	4 389	14 319	0.067	6.5	237
	7 009	20 440	0.030	5.8	175

Network Comparison

	$\lvert V\rvert$	$\lvert E\rvert$	clust. coeff.	avg. degree	max. degree
	4 389	14 319	0.067	6.5	237
	7 009	20 440	0.030	5.8	175



Simulations: Robustness of Algorithm



Conclusion & Outlook

Color-coding, with some algorithm engineering, is a practical and reliable method for finding signaling pathways in protein interaction networks.

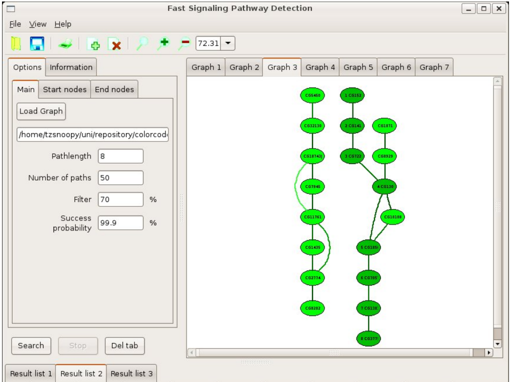
Conclusion & Outlook

Color-coding, with some algorithm engineering, is a practical and reliable method for finding signaling pathways in protein interaction networks.

Future work:

- Pathway queries
- Richer motifs (cycles, trees, ...)
- Derandomization

Graphical User Interface (upcoming)



	Weight	Prot 1	Prot 2	Prot 3	Prot 4	Prot 5	Prot 6	Prot 7	Prot 8	Selected
1	0.317429	CG6998	CG3227	CG5450	CG32130	CG18743|C(	CG7945	CG11761	CG5063	☐
2	0.323947	CG1871	CG8929	CG13030	CG10108	CG1856	CG7057	CG13811	CG3779	☑
3	0.339116	CG32130	CG18743|C(	CG7945	CG11761	CG17599	CG9740	CG4622	CG11454	☐
4	0.368402	CG5450	CG32130	CG18743|C(	CG7945	CG11761	CG1435	CG2774	CG8282	☑
5	0.373786	CG15283	CG14168	CG7224	CG13030	CG1856	CG7057	CG13811	CG3779	☑
6	0.391802	CG15468	CG14818	CG9951	CG6856	CG17599	CG9740	CG4622	CG11454	☐
7	0.416075	CG18591	CG16792	CG13277	CG6610	CG1249	CG8282	CG2774	CG1138	☐
8	0.433175	CG6425	CG5203	CG18743|C(	CG32130	CG5450	CG3183	CG6998	CG3227	☐