V10 Protein-Protein-Interaktionsnetzwerke

Rückblick (V7): Arten von PP-Interaktionen:

Homo-Oligomere vs. **Hetero-Oligomere**

Homo-Oligomere bestehen aus mehreren identischen Einheiten und werden z.B. von Ionen-Kanälen oder Rezeptoren gebildet

Stabile vs. **Transiente** Komplexe

Stabil: Ribosom, RNA-Polymerase, …

Transient: Redox-Partner, Signaltransduktion

Obligate vs. **Nicht-obligate** Komplexe:

obligat (-> obligatorisch): Komponenten liegen in der Zelle nur als Komplex vor

nicht-obligat: Komponenten existieren in der Zelle ebenfalls im freien Zustand (z.B. Antikörper)

Kovalente vs. **Nicht-kovalente** Komplexe

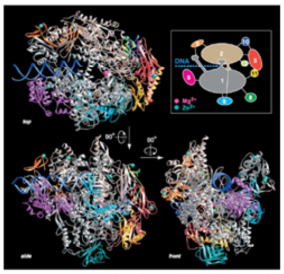
Kovalente: z.B. Ubiquitin-modifizierte Proteine

Nicht-kovalent: ist der übliche Fall

In Vorlesung 10 werden wir uns mit Proteinkomplexen und Proteininteraktionsnetzwerken beschäftigen.

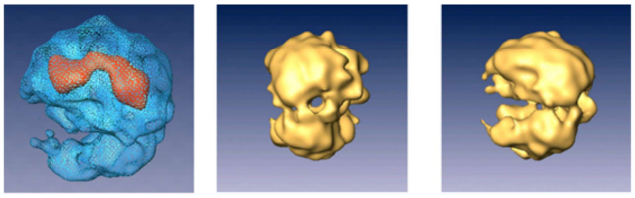
Beispiele für wichtige Proteinkomplexe

RNA Polymerase II



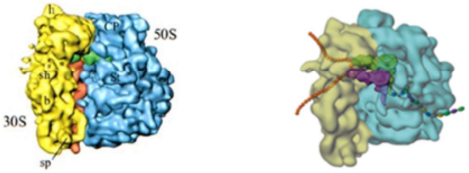
Cramer *et al.*, Science 288, 640 (2000)

Spleißosom



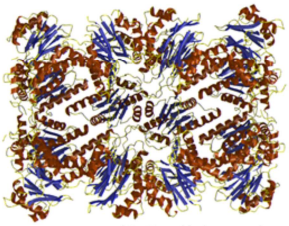
http://www.weizmann.ac.il/

Ribosom



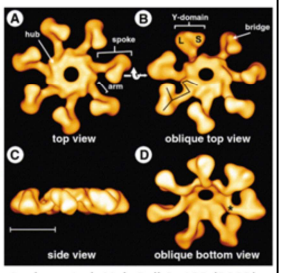
http://www.millerandlevine.com/chapter/12/cryo-em.html

Proteasome



http://www.biochem.mpg.de

Apoptosom

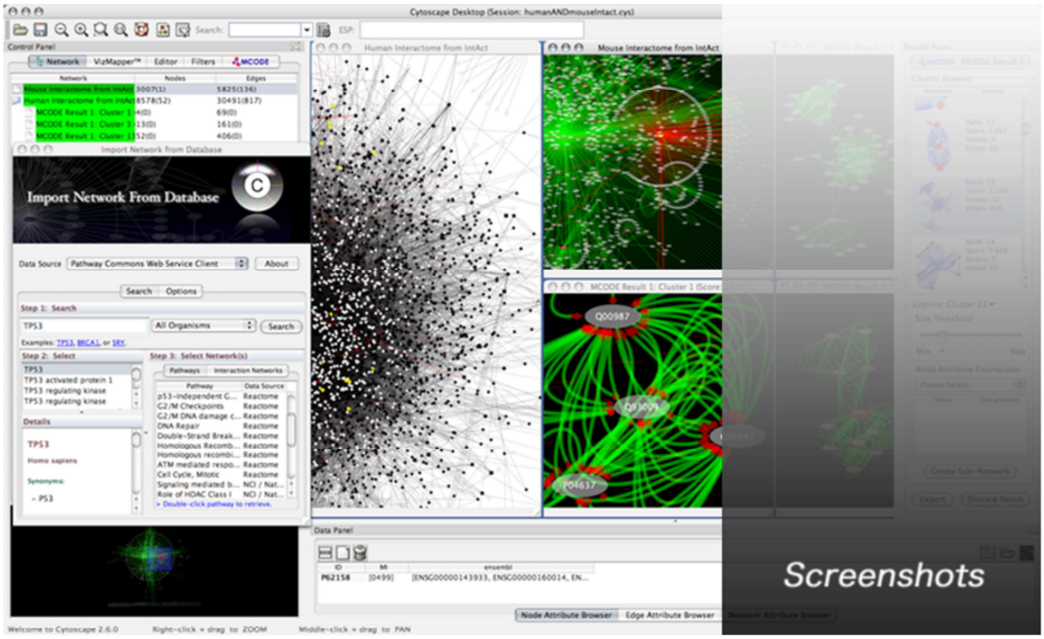


Acehan et al. Mol. Cell 9, 423 (2002)

Als Appetizer sind hier die dreidimensionalen Strukturen einiger sehr bekannter Proteinkomplexe gezeigt. RNA Polymerase ist ein Komplex aus 10 einzelnen Proteinen. Das Spleißosom wird aus fünf verschiedenen kleinen nuklearen RNAs (snRNAs) und etwa 150 Proteinen zusammengesetzt. Es ändert jedoch während des Spleißprozesses seine Zusammensetzung und Konformation auf drastische Weise, siehe Video https://www.annualreviews.org/doi/suppl/10.1146/annurev-biochem-091719-064225

Das Ribosom enthält zwei unterschiedlich große Untereinheiten (50S und 30S). Bakterielle und eukaryontische Ribosomen enthalten einen gemeinsamen strukturellen Kern, der aus 34 konservierten Proteinen (15 in der kleinen Untereinheit, 19 in der großen) und etwa 4400 RNA Basen besteht. Spleißosom und Ribosom enthalten jeweils RNA-Moleküle, um damit Kontakte zur mRNA zu bilden. Das Proteasom ist für den Ubiquitin-gesteuerten Abbau von Proteinen verantwortlich. Es besteht aus über 60 Einzelproteinen. Das Apoptosom besitzt eine ungewöhnliche siebenzählige Symmetrie und bildet sich in der Zelle, wenn der Zelltod (Apoptose) eingeleitet wird.

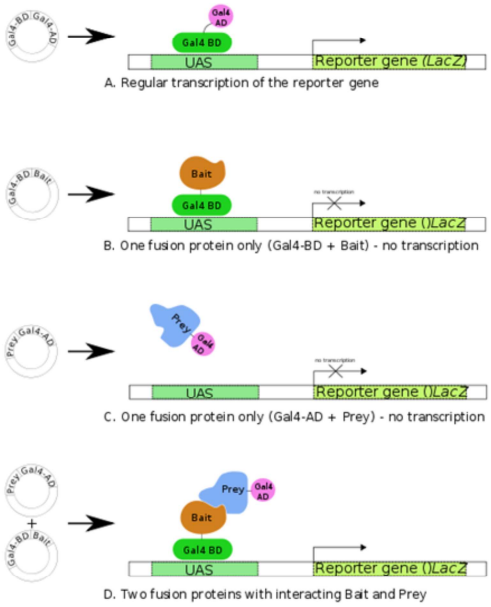
Proteininteraktionsnetzwerke



Proteininteraktionsnetzwerke sind Graphen-Darstellungen, welche Proteine (einzelne Punkte bzw. Knoten) physikalische Kontakte (Linien bzw. Verbindungen) miteinander eingehen können. Aufgrund der großen Anzahl an Knoten und Kanten bezeichnet man solche Netzwerke als „hairball monsters“. Eigentlich sieht man darin gar nichts. Wir kommen am Ende der heutigen Vorlesung auf solche Netzwerke zurück.

Detektiere PP-Interaktionen: Yeast Two-Hybrid Methode

Ziel: entdecke binäre PPIs zwischen einem "bait" Protein (*dt. Köder*) und einem "prey" Protein (*dt. Beutetier*), die "physikalisch", d.h. direkt miteinander wechselwirken.



Gegeben: Transkriptionsfaktor, der ein Reporter-Gen reguliert, besteht aus einer DNA-bindenden Domäne (BD) und einer Aktivatordomäne (AD)

Unterbreche kovalente Verbindung BD-AD; verbinde bait (orange) mit BD und prey (grün) mit AD → Expression findet nur statt, wenn bait:prey-Komplex gebildet wird.

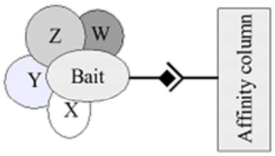
Man kann das normale Reporter-Gen (hier lacZ) auch durch ein GFP-Gen ersetzen. Dann kann man erfolgte Transkription, d.h. Bindung von bait:prey-Komplex, als Fluoreszenz detektieren.

10. Vorlesung WS 2020/21 www.wikipedia.org Softwarewerkzeuge 4

Die Hefe Zwei-Hybrid-Methode (Y2H) ist eine experimentelle Technik um direkte Proteininteraktionen experimentell zu detektieren. Die Methode wurde 1989 erstmalig vorgestellt (https://pubmed.ncbi.nlm.nih.gov/2547163/) und im Jahr 2000 (https://www.nature.com/articles/35001009) erstmalig als Hochdurchsatzmethode zur Detektion von etwa 1000 Interaktionen in Hefe eingesetzt, die 1000 der 6000 Hefe-Proteine involvierten. Mittlerweile schätzt man, dass es in Hefe etwa 90.000 Proteininteraktionen gibt (siehe Folie 10), d.h. 30 pro Protein, da ja an jeder Interaktion zwei Proteine beteiligt sind.

Tandem affinity purification (also „pull-down“)

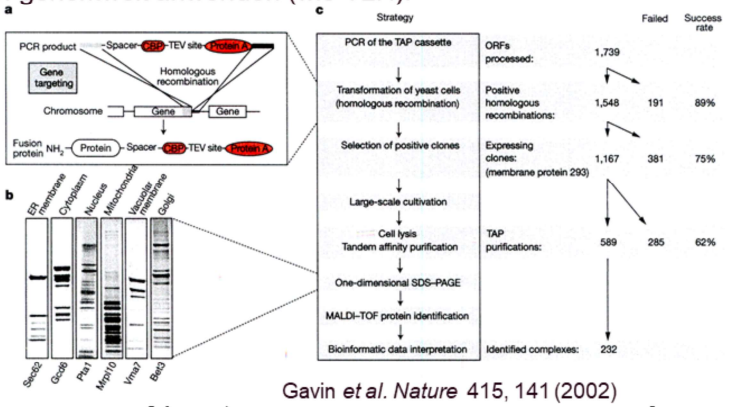
Die Yeast 2-Hybrid-Methode kann nur binäre Komplexe identifizieren.
In der **Affinitäts-Aufreinigung** wird ein bestimmtes Protein (bait) mit einem molekularen Label verbunden (dunkle Route in Abb.) um dessen Aufreinigung zu erleichtern.
Das so “ge-taggte” Protein wird dann in einer Affinitätssäule aus der Lösung herausgefischt, zusammen mit allen interagierenden Partnern (W, X, Y, Z).
Diese Strategie kann man genomweit anwenden (wie Y2H).



Hier gezeigt:
Anwendung für
S. cereviseae.

Identifiziere Proteine in Gelbanden mittels Massenspektrometrie.

Label unten: bait-Protein
Jede Bahn entspricht einem Komplex



Gavin *et al. Nature* 415, 141 (2002)

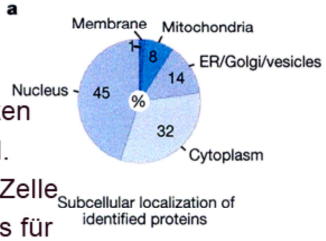
Bei der Tandemaffinitätsaufreinigung bindet man in einer Affinitätssäule Proteinkomplexe an das Trägermaterial. Man bringt zunächst an ein Protein der Zelle ein „tag“-Protein wie etwa Protein A an. In der Säule bindet dieses dann zusammen mit allen weiteren Proteinen, die evtl. daran gebunden sind, an einen dort angebrachten IgG-Antikörper. Mittlerweile werden viele unterschiedliche tags verwendet. Anschließend wäscht man den gebundenen Komplex aus der Säule heraus, trennt seine Komponenten auf einem Gel auf und bestimmt deren Identität mit Massenspektroskopie.

TAP-Analyse für PP-Komplexe in *S. cereviseae*

Identifiziere Proteine durch die Massen ihrer Peptid-fragmente.

(a) listet die an Komplexen beteiligten Proteine bzgl. ihrer Lokalisation in der Zelle -> es scheint keinen Bias für bestimmte Kompartments zu geben.

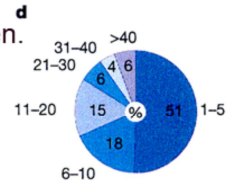
Allerdings findet man nur wenig Membran-proteine (Anteil sollte ca. 25% sein)



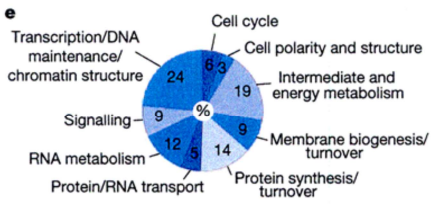
Subcellular localization of identified proteins

(d) die Hälfte aller PP-Komplexe hat 1-5 Mitglieder, die andere Hälfte ist grösser.

(e) PP-Komplexe sind an praktisch allen zellulären Prozessen beteiligt.



Number of proteins per complex



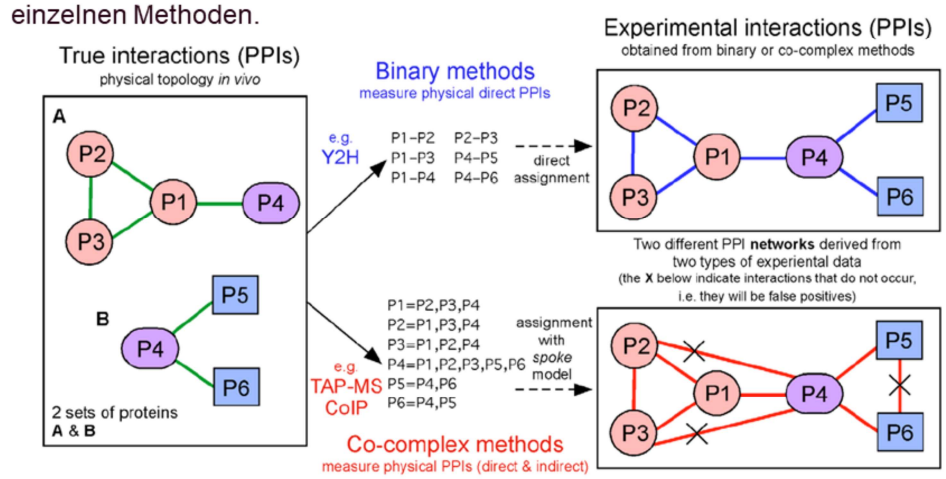
Distribution of complexes according to function

Gavin *et al. Nature* 415, 141 (2002)

Die TAP-Analyse für Hefe ergab 232 Proteinkomplexe. Abb. (a) zeigt, dass diese Komplexe aus verschiedenen Zellkompartementen stammen. Lediglich die Membranfraktion scheint unterrepräsentiert. Etwa 25-30% aller zellulären Proteine lokalisieren in der Membran und auch dort bilden sich zahlreiche Membranproteinkomplexe. Allerdings können diese Interaktionen durch die Durchführung der Affinitätsaufreinigung in der flüssigen Phase nur sehr bedingt detektiert werden.

Protein-Interaktionsnetzwerke

Unterschiedliche experimentelle Techniken messen verschiedene Eigenschaften von Proteinkomplexen. Das sind keine Messfehler, sondern Eigenheiten der einzelnen Methoden.



In der Probe gibt es zwei Arten von Komplexen, A und B

Ergebnis der Messung für diese Probe.

„Logische“ Interpretation der Messergebnisse.

De Las Rivas, PLOS Comp Biol. 6, e1000807 (2010)

Diese Folie illustriert an einem Beispiel, was man mit der Y2H-Methode (oben) und mit der TAP-MS (unten) detektieren würde und welche Schlüsse man aus diesen Ergebnisse ziehen würde. Im linken Beispiel befinden sich zwei Proteinkomplexe A und B in einer Probe. Y2H „vermischt“ die beiden Komplexe aufgrund der existierenden Kontakte zu einem einzigen großen Komplex. Y2H kann nicht auflösen, ob die beobachteten Kontakte zu einem einzelnen oder zu mehreren Komplexen gehören. Aus der TAP-MS enthält man keine Information, welche Einzelproteine direkt miteinander interagieren.

Globales Protein-Interaktionsnetzwerk in *S. cerevisiae*

Abb. zeigt das Protein–Protein Interaktions-netzwerk in *Saccharomyces cerevisiae*, basierend auf **yeast two-hybrid** Experimenten.

Jeder Knoten (Kreis): bestimmtes Protein
Verbindungen/Kanten: physikalische Interaktionen.

Das Netzwerk wird von einigen Knoten mit sehr vielen Verbindungen zusammengehalten. Diese nennt man **Hubs**.

Dieser größte Cluster enthält 78% aller Proteine der Zelle.

Rot: Gendeletion ist tödlich

Grün: Gendeletion ist nicht tödlich

Gelb: Effekt der Gendeletion ist unbekannt.

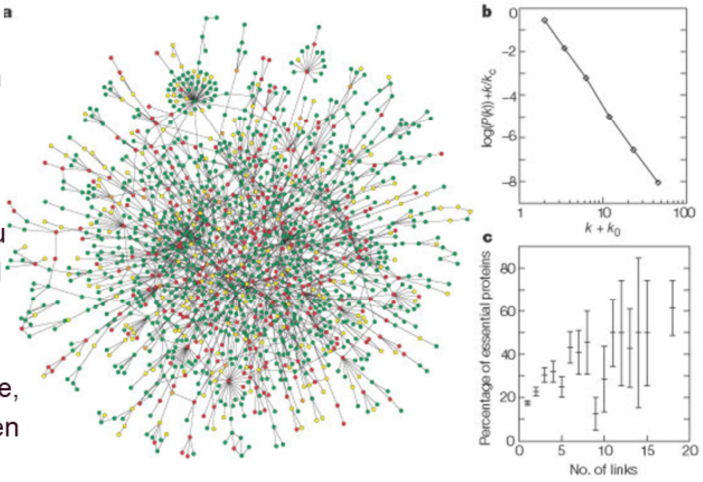
Barabasi & Oltvai, Nature Rev Gen 5, 101 (2004)

Hier interessierte man sich dafür, welche Positionen des Netzwerkes essentielle Proteine (rot) bzw. nicht-essentielle Proteine (grün) einnehmen.

Welche Proteindeletionen sind tödlich?

(b) Die Häufigkeit von Hub-Proteinen mit *k* Interaktionen nimmt nicht exponentiell ab (wie man dies in einem Zufallsnetzwerk erwartet), sondern etwa proportional zu $1 / k^2$ oder $1 / k^3$, ist also viel häufiger als zufällig erwartet.

(c) Die Deletion solcher Gene, die für Hub-Proteine mit vielen Links kodieren (und im Plot rechts liegen), ist eher **tödlich** für die Zelle als die Deletion von Proteinen, die mit wenigen anderen Proteinen interagieren (links im Plot).



„essentielle“ Proteine: Zelle ist nicht lebensfähig ohne diese Proteine

H. Jeong, S. P. Mason, A.-L. Barabási and Z. N. Oltvai
Nature ***411, 41-42***

In Abb. (b) ist die Häufigkeit von Proteinen mit unterschiedlich vielen Interaktionen gezeigt. Die Häufigkeit fällt mit etwa $1/k^3$ ab. Es gibt daher deutlich mehr „Hub“-Proteine mit sehr vielen Interaktionen als man in einem Zufallsnetzwerk erwarten würde. Solche Netzwerke, in denen die Häufigkeit mit einem Potenzgesetz abnimmt, nennt man „scale free“. Mehr zu solchen Netzwerken behandeln wir in der Vorlesung Bioinformatik 3.

In der Abbildung rechts unten sieht man, dass der Anteil essentieller Proteine unter den Hub-Proteinen bei etwa 60% liegt, im Vergleich zu 20% für Proteine mit sehr wenigen Interaktionen. Die Deletion eines solchen Hub-Proteins ist daher für die Hefe-Zelle mit recht hoher Wahrscheinlichkeit tödlich.

Wieviele Proteininteraktionen gibt es

S. cerevisiae

BioGrid (www.thebiogrid.org): 91,651 nicht-redundante physikalische Interaktionen von 6367 Hefe-Proteinen (August 2017).

Mentha (http://mentha.uniroma2.it/): 106,683 Interaktionen.

PrePPI (https://bhapp.c2b2.columbia.edu/PrePPI/).
(bioinformatisch integrativ kompilierter Datensatz):
60000 Interaktionen mit hoher Zuverlässigkeit

Mensch

Mentha: 277,371 physikalische Interaktionen zwischen 18,506 menschlichen Proteinen.

PrePPI: 1.35 Millionen vorhergesagte Interaktionen.
Für 127,000 davon gibt es experimentelle Bestätigung der direkten Interaktion.

Verschiedene Datenbanken enthalten etwas unterschiedliche Informationen für Hefe und Mensch. Im Mensch gibt es etwas dreimal mehr Interaktionen als in Hefe.

In-Silico Vorhersagemethode

Sequenz-basiert

Funktionelle Ähnlichkeit:

- Gen-Clustering
- Gen-Nachbarschaft

Interaktionen:

- Rosetta stone
- phylogenetisches Profiling
- Ko-Evolution

Struktur-basiert:

- interface propensities (V7)
- Protein-Protein Docking (V7)
- 3D-Simulationen (z.B. MD)

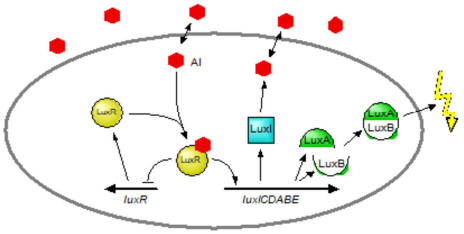
Auf den folgenden Folien betrachten wir einige einfache Bioinformatik-Methoden, mit denen man Protein-Interaktionen vorhersagen kann. Diese sind links aufgelistet. Rechts sind ein paar strukturbasierte Methoden aufgeführt, die wir bereits in Vorlesung 7 kennengelernt hatten.

Gen-Clustering

Idee: funktionell **verwandte** Proteine oder Teile eines Komplexes werden **gleichzeitig** exprimiert

Suche nach Genen mit einem **gemeinsamen Promoter**
→ wenn aktiviert, werden alle gemeinsam als ein *Operon* transkribiert.

Beispiel:
Biolumineszenz in *V. fischeri* wird durch Quorum sensing reguliert
→ 3 Proteine I, AB, CDE sind dafür verantwortlich.
Sie sind als 1 Operon namens *luxICDABE* organisiert.



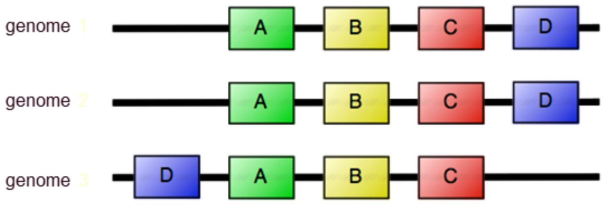
10. Vorlesung WS 2020/21 Softwarewerkzeuge 12

Die Gene, die für funktionell eng verwandte Proteine kodieren, haben in Bakterien teilweise einen gemeinsamen Promoter und werden eins nach dem anderen exprimiert. Dies bezeichnet man als **Operon**. Diese Organisation gibt es nur in Prokaryoten. Auch in Eukaryonten liegen funktionell verwandte Proteine jedoch oft benachbart auf dem Genom, haben aber eigene Promoterabschnitte.

Gen-Nachbarschaft

Hypothese: funktionell **verwandte** Gene werden **gemeinsam exprimiert**

"funktionell verwandt" heißt gleicher {Komplex | Pfad | Funktion | …}

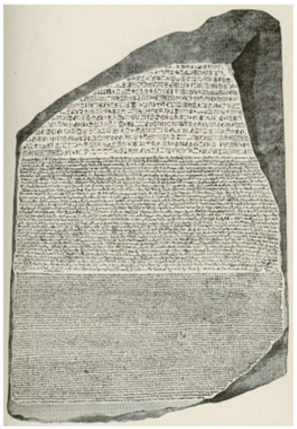


→ Suche nach **ähnlicher Anordnung** der verwandten Gene in **verschiedenen Organismen**

(<=> Gen-Clustering: in einem Organismus, Promoter müssen bekannt sein

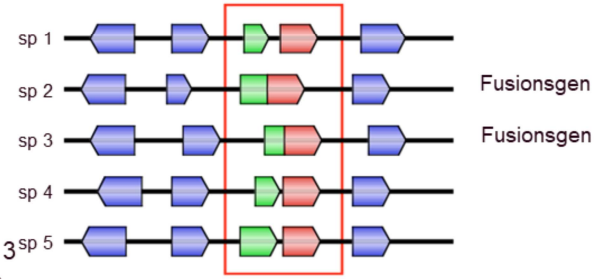
Das Prinzip der Gen-Nachbarschaft kann, wie erwähnt, auf funktionelle Verwandtschaft von Proteinen hindeuten. Man kann daher gezielt nach sequentiell angereihten Genen suchen, deren Anordnung in etlichen verschiedenen Organismen konserviert ist. Diese Konservierung deutet dann wiederum auf deren gemeinsame funktionelle Rolle hin.

Rosetta Stein Methode



Mehrsprachige Stele aus 196 v.Chr., wurde 1799 gefunden.
Auf dem Stein steht derselbe Text in 3 Sprachen: Hieroglyphen, demotische Schrift, griechisch
→ Schlüssel um Bedeutung der Hieroglyphen zu entschlüsseln

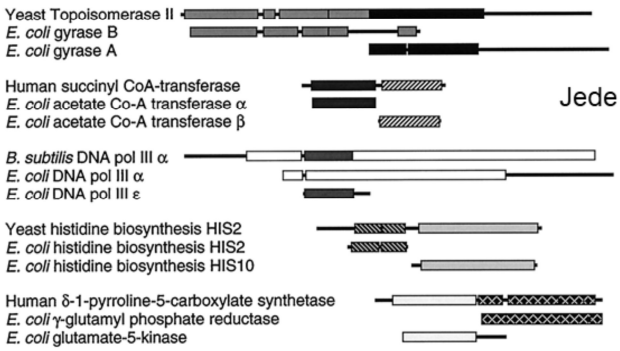
Idee: finde homologe Gene ("**Worte**") in den Genomen verschiedener Organismen ("**Texte**")
- Überprüfe, ob ein Organismus ein Fusions-Gen enthält
→ Kann darauf hindeuten, dass die beiden Proteine einen Komplex bilden



Enright, Ouzounis (2001):
40000 vorhergesagte paarweise Interaktionen in 23 Spezies

Wenn bei zwei interagierenden Proteinen der C-Terminus des einen Proteins im Komplex räumlich benachbart zum N-Terminus des zweiten Proteins liegt, ist eine Fusion der beiden Gene auf genomischer Ebene möglich.

Rosetta Stein Methode



Jede Box symbolisiert ein Protein.

5 Beispiele von *E. coli*-Proteinen, deren Interaktion vorhergesagt wurde.

Da in einem anderen Organismus (erste Reihe) ein Fusionsprotein existiert, ist es wahrscheinlich, dass das zweite und dritte Protein aus *E. coli* interagieren.

Die ersten 3 Interaktionen waren aus dem Experiment bereits bekannt.

Marcotte et al. Science 285, 751 (1999)

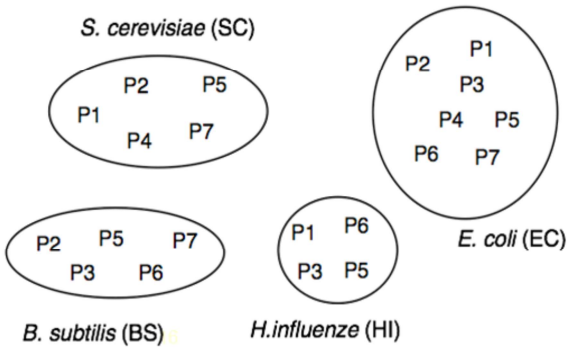
In https://science.sciencemag.org/content/285/5428/751.full
wurde die Rosetta-Stone-Methode zum ersten Mal vorgestellt.

Für diese Rosetta-Stone-Methode wurde sogar ein Patent in den USA erteilt: https://www.osti.gov/doepatents/biblio/874814

Phylogenetisches Profiling

Idee: entweder **alle** oder **kein** Protein eines Komplexes sollten in einem Organismus vorkommen

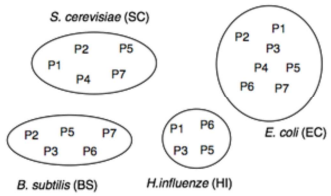
→ Vergleiche Vorkommen homologer Proteine zwischen Spezies (z.B. mit Sequenzalignment)



Hinter der Methode „phylogenetisches Profiling“ steht die Idee, dass zwei interagierende Proteine natürlich beide im Genom des Organismus vorkommen müssen. Wenn eines nicht vorhanden ist, kann die Interaktion nicht stattfinden, obwohl das andere Gen trotzdem im Genom vorhanden sein könnte. Wenn man jedoch eine Statistik über eine große Anzahl an Organismen macht, stellt man fest, dass von zwei interagierenden Proteinen oft entweder beide vorhanden sind oder keines.

Diese Beobachtung kann man natürlich auch umdrehen. Man sucht nach Proteinpaaren, deren Auftreten korreliert ist und nutzt dies als Vorhersage von möglichen Interaktionen.

Distanzen in Phylogenetischem Profiling

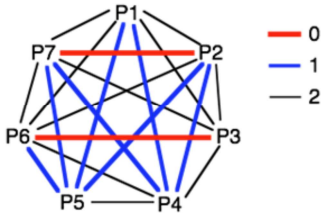


Kodiere Vorkommen/Abwesenheit

	EC	SC	BS	HI
P1	1	1	0	1
P2	1	1	1	0
P3	1	0	1	1
P4	1	1	0	0
P5	1	1	1	1
P6	1	0	1	1
P7	1	1	1	0

Hamming-Distanz zwischen Spezies: Anzahl an unterschiedlichem Vorkommen

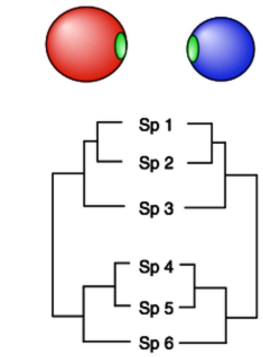
	P1	P2	P3	P4	P5	P6	P7
P1	0	2	2	1	1	2	2
P2		0	2	1	1	2	**0**
P3			0	3	1	**0**	2
P4				0	2	3	1
P5					0	1	1
P6						0	2
P7							0

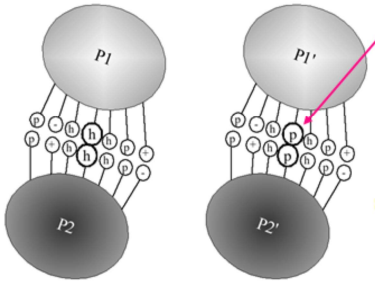


Paare mit ähnlichem Vorkommen sind: P2-P7 und P3-P6
Dies sind Kandidaten für Protein-Interaktionen.

In diesem einfachen Beispiel betrachten wir (als Beispiel) den eukaryontischen Organismus Hefe (*S. cerevisae*) und drei Bakterienstämme. Das Vorkommen/Nichtvorkommen der sieben Proteine P1 bis P7 (bzw. ihrer Gene im Genom) wird mit 1 und 0 symbolisiert.

Dann berechnen wir zwischen zwei Proteinen (Reihen) die Summe ihrer Unterschiede. Dies bezeichnet man als Hamming-Distanz. Z.B. kommen P2 und P7 gemeinsam in EC, SC und BS vor und fehlen beide in HI. Deshalb haben sie die Hamming-Distanz 0.





Ko-Evolution

Bindungsschnittstellen von Komplexen sind nur **leicht stärker konserviert** als die restliche Proteinoberfläche.

Idee von Pazos & Valencia (1997):
Falls an einer Schnittstelle eine Mutation auftritt, die den Charakter der Aminosäure ändert (z.B. hydrophob/hydrophob in P1/P2 -> polar/polar in P1'/P2'),
sollten an dem anderen Interface korrespondierende Mutationen an den Positionen auftreten, die mit der ersten Aminosäure Kontakte bilden.

Die Suche nach solchen korrelierten Mutationen kann dabei helfen, Bindungskandidaten zu finden.

10. Vorlesung WS 2020/21 Softwarewerkzeuge 18

Eine ähnliche Idee wie beim phylogenetischen Profiling kann man auch auf einzelne Aminosäurepositionen anwenden. Zwei Aminosäuren, die sich an einer Bindungsschnittstelle gegenüber liegen, haben vermutlich einen ähnlichen physikochemischen Charakter (hydrophob bzw. polar). Allerdings könnten in einem anderen Organismus beide Aminosäuren zusammen von hydrophob nach polar mutiert sein, was ebenfalls eine günstige Wechselwirkung bewirken würde. Über viele Organismen gesehen, würde man daher eine Korrelation der Hydrophobizitäts-Werte von direkt interagierenden Aminosäurepositionen erwarten.

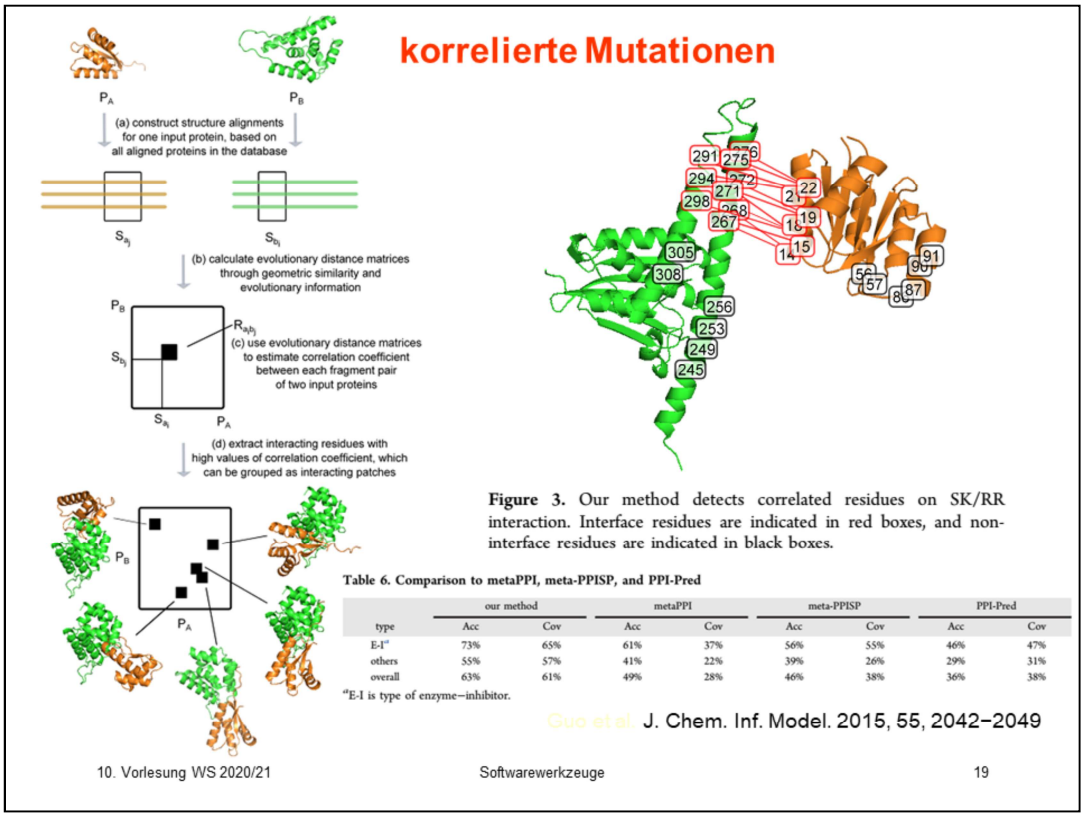


Figure 3. Our method detects correlated residues on SK/RR interaction. Interface residues are indicated in red boxes, and non-interface residues are indicated in black boxes.

Table 6. Comparison to metaPPI, meta-PPISP, and PPI-Pred

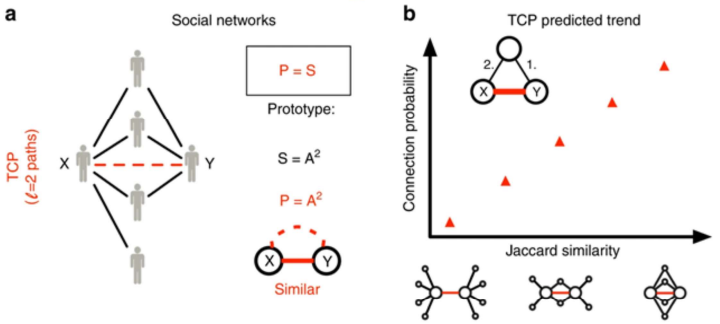
	our method		metaPPI		meta-PPISP		PPI-Pred	
type	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov
E-I[a]	73%	65%	61%	37%	56%	55%	46%	47%
others	55%	57%	41%	22%	39%	26%	29%	31%
overall	63%	61%	49%	28%	46%	38%	36%	38%

[a]E-I is type of enzyme−inhibitor.

Guo et al. J. Chem. Inf. Model. 2015, 55, 2042−2049

Man kann gezielt in multiplen Sequenzalignments von zwei Proteinen nach solchen Korrelationen zwischen Aminosäurepositionen suchen. Wichtig ist hierbei, dass die Organismen in derselben Reihenfolge von oben nach unten angeordnet sind, da man ja Änderungen aufspüren möchte, die jeweils gemeinsam auftreten. Diese Methode ist sehr erfolgreich um interagierende Aminosäurepaare aufzuspüren, siehe die Abb. rechts oben. Starke Korrelationen zwischen Aminosäurepaaren gibt es in diesen beiden Proteinsequenzen nur am Bindungsinterface. Die beiden Interaktionspartner wurden hier zur besseren Visualisierung leicht auseinandergezogen.

Link-Vorhersage basierend auf Netzwerkdaten



Kovács, ... Vidal & Barabási
Nature Commun. 10, 1240 (2019)

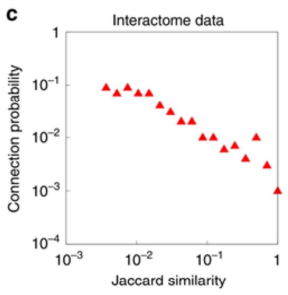
(a) In sozialen Netzwerken impliziert eine große Anzahl an gemeinsamen Freunden eine hohe W'keit, dass 2 Leute Freunde werden (rote Verbindung zwischen Knoten X und Y). Dies nennt man **Triadic Closure Principle (TCP)**.

TCP sagt basierend auf einer Knotenähnlichkeit (*S*) Verbindungen *(P)* voraus. Ein Maß für die Ähnlichkeit ist z.B. die Anzahl an gemeinsamen Nachbarn zwischen jedem Knotenpaar (A^2).

(b) Übertragen auf PPI-Netzwerke würde man erwarten, dass Proteinpaare mit hoher Jaccard-Ähnlichkeit ebenfalls miteinander interagieren.

Man kann Interaktionen auch aufgrund der Position von Proteinen in Proteininteraktionsnetzwerken vorhersagen. In sozialen Netzwerken (links) ist die Anzahl an gemeinsamen Freunden ein gutes Indiz dafür, ob zwei Individuen X und Y ebenfalls Freunde werden könnten. Gilt dies auch für Proteine? Dies ist rechts symbolisiert. Dies würde bedeuten, dass 2 Proteine, die an mehrere gemeinsame Bindungspartner binden können, ebenfalls aneinander binden.

TCP trifft nicht auf PPI-Netzwerke zu



Kovács, ... Vidal & Barabási
Nature Commun. 10, 1240 (2019)

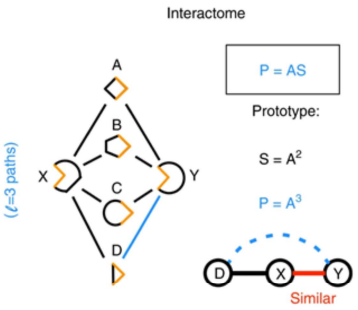
Jaccard-Koeffizient:
$J = |N_X \cap N_Y| / |N_X \cup N_Y|$,
wobei N_X and N_Y die Anzahl an Interaktionspartnern von X und Y sind.

Allerdings beobachten Kovács et al. in einem sehr zuverlässigen PPI-Netzwerk für den Mensch (HI-II-14) quasi das Gegenteil:

Proteinpaare mit hoher Jaccard-Ähnlichkeit haben eine geringere W'keit, miteinander zu interagíeren.

Die Anzahl an gemeinsamen Bindungspartner misst man mit dem Jaccard-Koeffizienten, d.h. dem Quotient aus der gemeinsamen Schnittmenge der Bindungspartner von X und Y geteilt durch deren Vereinigungsmenge. Der in der Abbildung tatsächlich beobachtete Trend ist jedoch abnehmend, nicht linear ansteigend, wie angenommen. Was ist der Grund hierfür?

PPIs involvieren Bindungsschnittstellen



PPIs benötigen meist komplementäre Schnittstellen. Deshalb teilen sich zwei Proteine, X und Y, mit ähnlichen Schnittstellen oft viele Bindungspartner.

Allerdings garantiert eine gemeinsame Schnittstelle nicht, dass X und Y direkt miteinander interagieren.

Stattdessen könnte ein weiterer Interaktionspartner von X (Protein D) ebenfalls mit Protein Y interagieren (**blaue Verbindung**).

Solche Kanten können durch Pfade der Länge 3 vorhergesagt werden (L3).

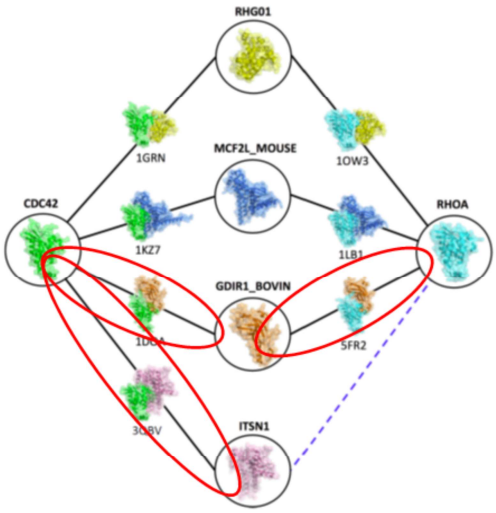
Kovács, ... Vidal & Barabási
Nature Commun. 10, 1240 (2019)

L3 identifiziert ähnliche Knoten zu den bekannten Partnern ($P = AS$), und geht dabei einen Schritt weiter als das Ähnlichkeitsmaß TCP.

Der Grund ist folgender: wenn X an Z binden kann und ebenfalls Y an Z binden kann, sind X und Y jeweils komplementär zu Z, jedoch anscheinend nicht zueinander. Z kann entweder X oder Y an seiner Bindungsschnittstelle binden (falls X und Y nicht auf verschiedenen Seiten an Z binden, was nicht sehr häufig der Fall ist).

Gut funktioniert jedoch eine andere Vorhersage: wenn X an D bindet und X und Y viele Bindungspartner gemeinsam haben, dann könnte Y ebenfalls an D binden. Solch eine Interaktion könnte man dadurch aufspüren, dass es 3 Pfade mit 3 Kanten zwischen Y und D gibt: Y – A– X – D, Y – B – X – D und Y – C – X – D.

Structurelle Veranschaulichung von L3



Kovács, ... Vidal & Barabási
Nature Commun. 10, 1240 (2019)

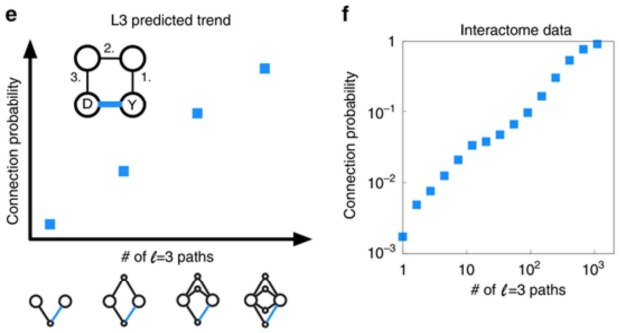
Gezeigt sind PDB-Strukturen für zwei menschliche Proteine, CDC42 und RHOA, die mit manchen Interaktionspartner durch das gleiche, gemeinsame Interface wechselwirken.

CDC42 und RHOA wechselwirken nicht miteinander. Sie könnten aber weitere gemeinsame Interaktionspartner haben, die an das gemeinsame Interface binden.

Z.B. deutet die **blau** gestrichelte Kante eine mögliche Interaktion zwischen ITSN1 und RHOA an.
Es existieren eine große Anzahl an Pfaden der Länge l = 3 in dem PPI-Netzwerk zwischen ihnen. Hier gezeigt sind 3 Pfade.

Hier ist ein tatsächliches Beispiel für solche L3-Pfade gezeigt.

L3 trifft auf PPI-Netzwerke zu!



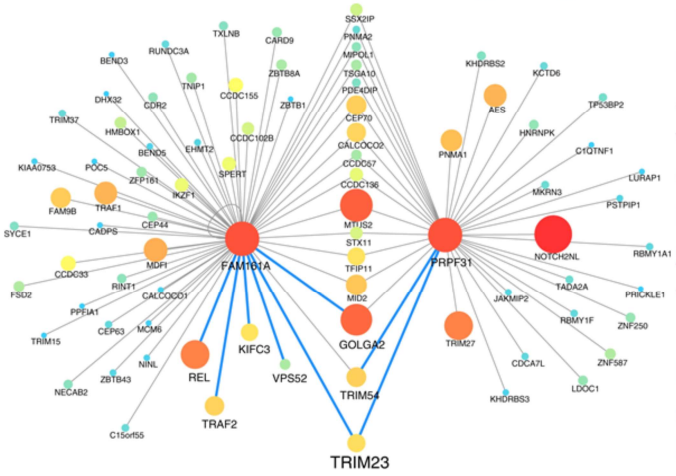
e Selbst ohne existierende Strukturinformation kann man erwarten, dass 2 Proteine Y und D miteinander interagieren, wenn sie durch mehrere Pfade der Länge ℓ = 3 im PPI-Netzwerk verbunden sind. (L3).

f Für das Benchmark PPI-Netzwerk HI-II-14 für menschliche Proteine wurde eine positive Korrelation entsprechend **e** beobachtet.

Kovács, ... Vidal & Barabási
Nature Commun. 10, 1240 (2019)

Die rechte Abbildung zeigt, dass die Interaktionswahrscheinlichkeit zwischen Y und D logarithmisch mit der Anzahl an l = 3-Pfaden zwischen Y und D ansteigt und im Fall von 1000 Pfaden den Wert 1 erreicht.

Durch L3 vorhergesagte Interaktion



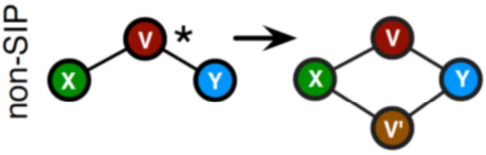
Kovács, ... Vidal & Barabási
Nature Commun. 10, 1240 (2019)

FAM161A and PRPF31 sind 2 Proteine, die mit der Krankheit *retinitis pigmentosa* verknüpft werden.
Gezeigt sind alle bekannten Interaktionspartner (grau), zusammen mit den durch L3 vorhergesagten (blau).

Die stärkste durch L3 vorhergesagte Interaktion verbindet FAM161A mit GOLGA2.
Die beiden Proteine haben keine gemeinsamen Interaktionspartner.
Knotengrösse und –farbe symbolisieren die Anzahl an Interaktionen.

Hier wurde die Interaktion zwischen FAM161A und GOLGA2 durch die Analyse der L3-Pfade vorhergesagt und später experimentell bestätigt.

Verbindung zur Evolution



Genduplikation ist ein Schlüsselmechanismus der Evolution, durch den neue Proteine entstehen können.

Wenn Protein V dupliziert wird (bzw. das dafür kodierende Gen), wird der duplizierte Knoten (V') zumindest anfangs die Interaktionen des Originalproteins V mit X und Y behalten.

Dies kann den Erfolg von L3 teilweise erklären.

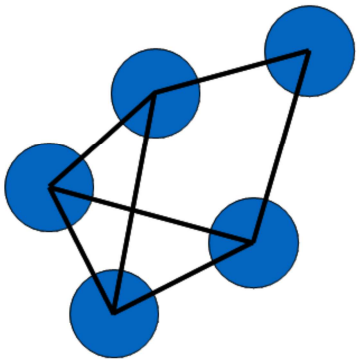
Kovács, ... Vidal & Barabási
Nature Commun. 10, 1240 (2019)

Wie könnten solche Proteinpaare entstanden sein? Ein möglicher Mechanismus ist die Genduplikation.

Spezifische PP-Interaktionsnetzwerke für bestimmte Bedingungen

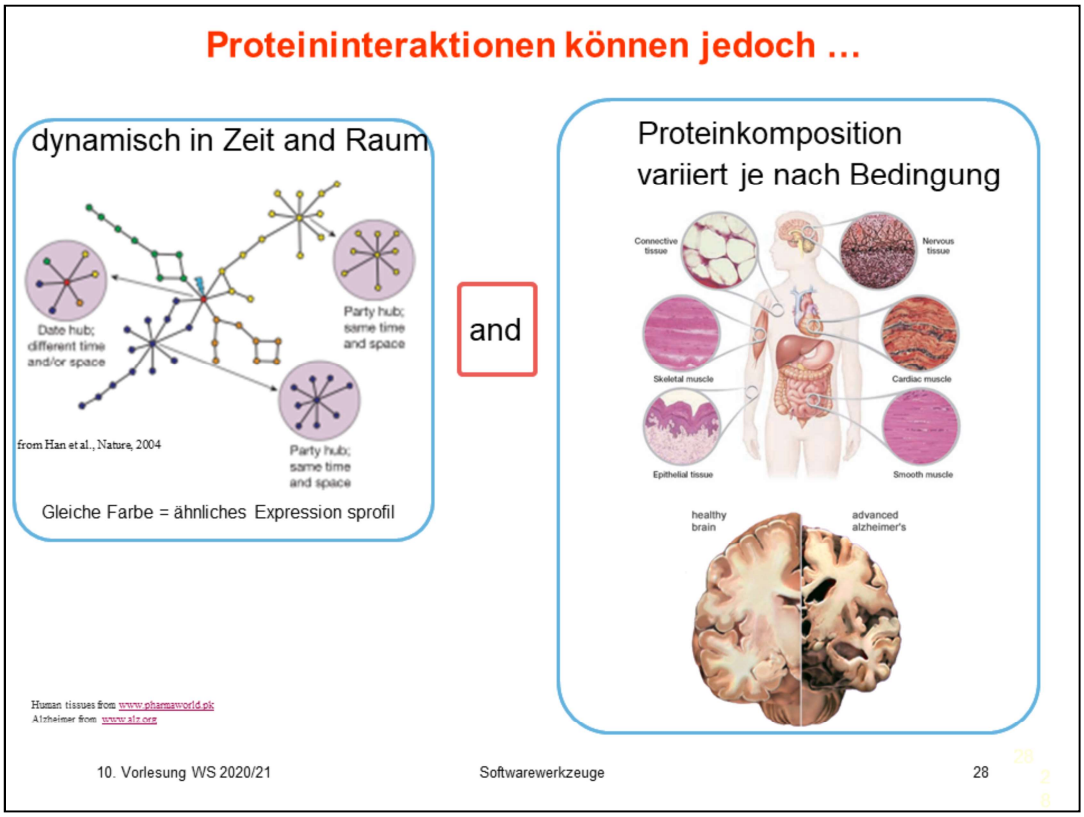
Modell: Brustkrebs (gute Datenlage)

Abb. zeigt das gesamte PPI-Netzwerk, z.B. für Mensch = Sammlung paarweiser Interaktionen aus **verschiedenen** Experimenten

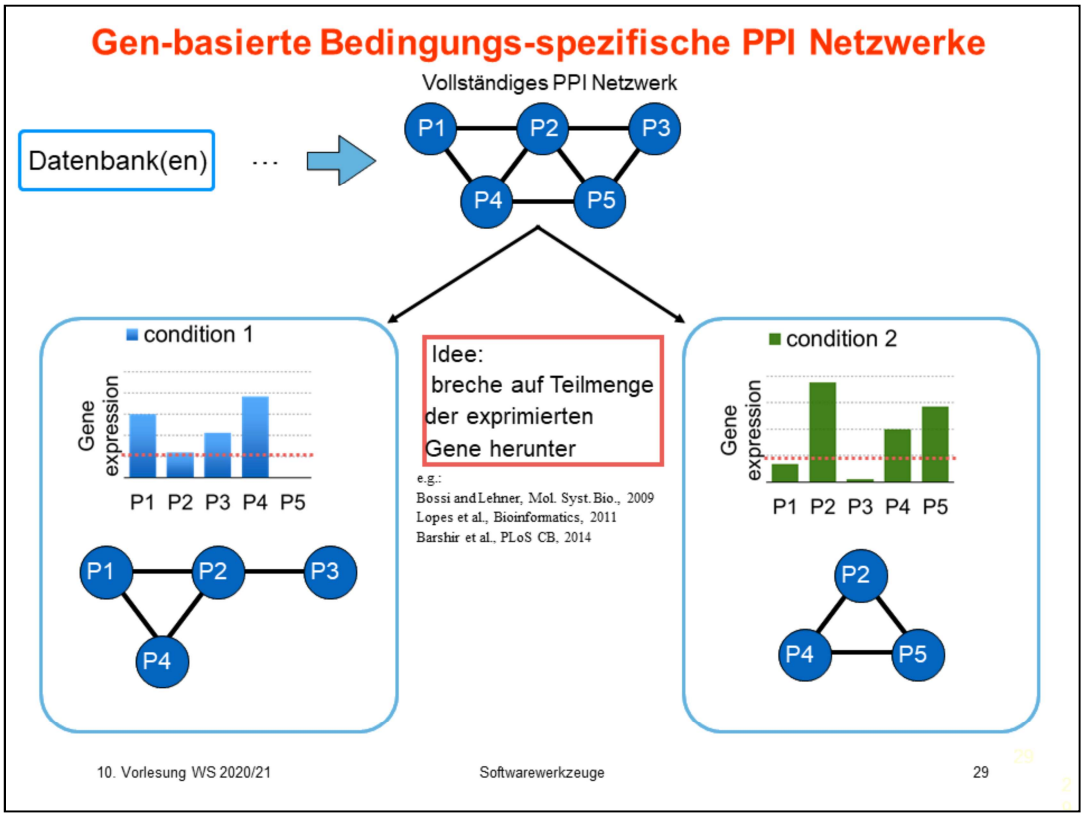


Will, Helms, Bioinformatics, 47, 219 (2015)
doi: 10.1093/bioinformatics/btv620

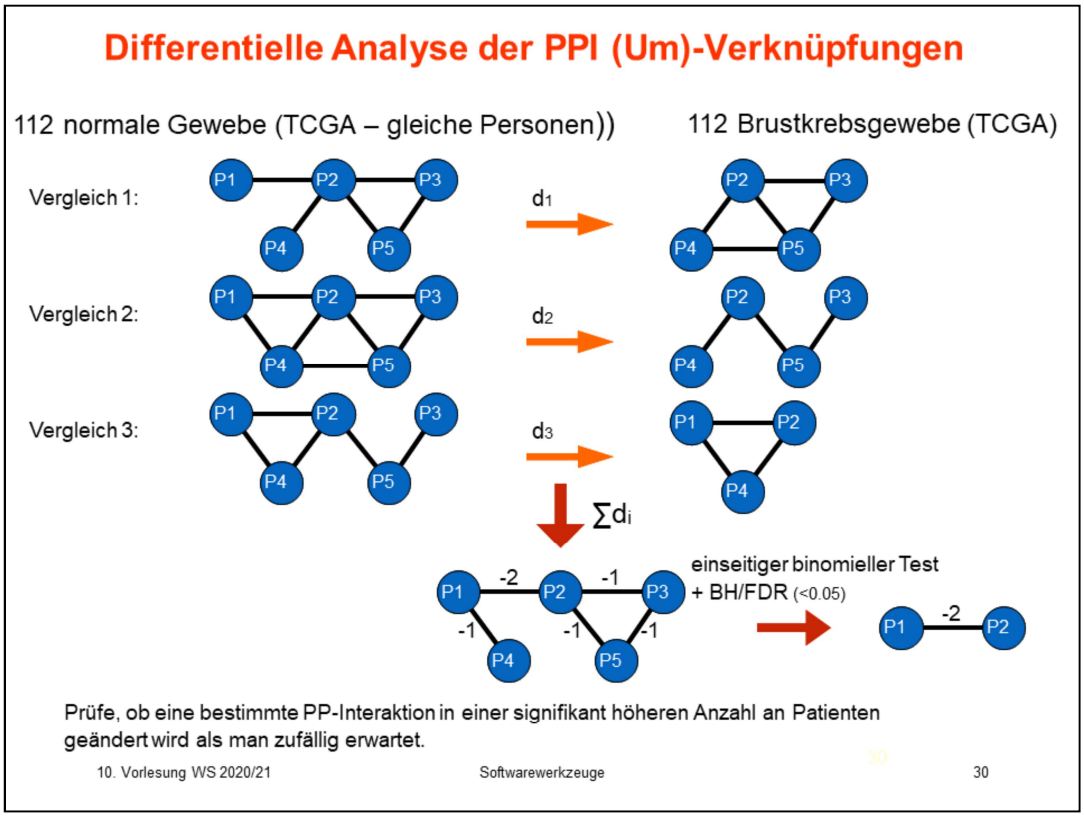
Proteininteraktionsnetzwerke bestehen aus paarweisen Interaktionen, die in zum Teil unabhängigen Experimenten detektiert wurden. Alle Interaktionen werden in einem statischen Netzwerk/Graphen angeordnet.



In Wirklichkeit sind diese Netzwerke aber sehr dynamisch, da einzelne Proteine zu unterschiedlichen Zeiten oder nur in unterschiedlichen Bedingungen für einen bestimmten Zellzustand oder in einem Krankheits-Phänotyp synthetisiert werden. Die Netzwerkdarstellung bildet dies aber nicht ab. Um diese Effekte einzubauen, kann man auf Expressionsdaten zurückgreifen.



Gewöhnlich filtert man dabei das Netzwerk, so dass nur die in einer bestimmten Bedingung exprimierten Gene übrigbleiben. Eine Interaktion kann natürlich nur existieren, wenn beide Interaktionspartner exprimiert werden.



Wir interessieren uns gewöhnlich dafür, welche Unterschiede zwischen zwei Zellzuständen existieren, z.B. in gesundem Gewebe und in Tumorgewebe. In einem bestimmten Patienten können ganz individuelle Verschaltungen der Protein-Interaktionen vorliegen, d.h. bestimmte Interaktionen existieren entweder nur im Tumor oder nur im gesunden Gewebe.

Hier führten wir solch einen Vergleich für 112 Brustkrebs-Patienten aus dem TCGA-Datensatz durch. Links ist immer das gefilterte Interaktionsnetzwerk für sein gesundes Brustgewebe gezeigt, rechts das für das Brustkrebsgewebe. Mit d bewerten wir für jede Kante, ob sie entweder nur im Tumor existiert (+1) oder nur im gesunden Gewebe (-1). Dann zählen wir alle d-Werte der Kanten für die 112 Patienten zusammen und vergleichen sie mit den zufällig erwarteten Verschaltungen. Nur diejenigen werden übrig behalten, die signifikant häufiger ($p < 0.05$) verschaltet wurden als zufällig erwartet wird.

Binomieller Verteilung / Test

Die diskrete Wahrscheinlichkeitsverteilung mit der Wahrscheinlichkeitsfunktion

$$B(k \mid p, n) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, \ldots, n$$

heißt die *Binomialverteilung* zu den Parametern *n* (Anzahl der Versuche) und $p \in [0, 1]$ (der *Erfolgs-* oder *Trefferwahrscheinlichkeit).*

Ein **Binomialtest** ist ein statistischer Test, bei dem die Testgröße binomialverteilt ist. Er wird verwendet um Hypothesen über Merkmale zu prüfen, die genau zwei Ausprägungen annehmen können.

In Fall der PP-Interaktionen kann eine Interaktion existieren oder nicht.

Man erzeugt also für eine genauso große Anzahl von 112 PP-Netzwerken dieselbe Anzahl von etwa 10.000 zufälligen Verschaltungsänderungen
Dann zählt man ab, wie häufig eine bestimmte Interaktion Pi-Pj zufällig „rewired“ wird und vergleicht dies mit der in Patienten beobachten Anzahl.
Daraus erhält man einen p-Wert für diese Interaktion Pi-Pj.

Die zufällig erwartete Anzahl an Verschaltungen erhält man durch ein (n=112 mal durchgeführtes) Experiment, bei dem man 10.000 Kanten des Netzwerks zufällig auswählt und verschaltet, d.h. entfernt oder hinzufügt.

Rewiring von PPIs in Brustkrebs vs. gesundem Gewebe

	GENE
avg. number of proteins (normal)	$12,678 \pm 223$
avg. number of proteins (tumor)	$12,528 \pm 206$
avg. number of interactions (normal)	$134,348 \pm 2,387$
avg. number of interactions (tumor)	$133,128 \pm 2,144$
P_{rew}	0.067 ± 0.016
significantly rewired interactions	9,754

Im Mittel liegen 12.500 – 12.600 Proteine vor.

Die Standardabweichung drückt Unterschiede zwischen einzelnen Patienten aus.

Anhand der bekannten Interaktionsdaten erwartet man zwischen diesen Proteinen etwa 134.000 PP-Interaktionen

-> etwa 10.000 dieser PP-Interaktionen sind in Krebs-Gewebe signifikant anders verschaltet als in gesundem Gewebe.

Im Mittel enthalten die Interaktionsnetzwerke in gesundem Gewebe und in Tumorgewebe beide jeweils etwa 12.500 Proteine und etwa 134.000 Interaktionen. Allerdings sind im Durchschnitt 9754 Interaktionen „verschaltet“. Dies ist eine riesige Anzahl und resultiert daraus, dass vermutlich ca. 2000 Gene zwischen beiden Geweben differentiell exprimiert sind.

Rewired PPIs sind mit Krebs-Merkmalen assoziiert

	GENE
rewired interactions	9,754
participation in any hallmark term	7,028
fraction in any hallmark term	0.721
Resisting Cell Death	4,064 (0.417)
Activating Invasion and Metastasis	2,244 (0.230)
Sustaining Proliferative Signaling	3,964 (0.406)
Inducing Angiogenesis	169 (0.017)
Tumor-Promoting Inflammation	516 (0.053)
Genome Instability and Mutation	1,362 (0.140)
Enabling Replicative Immortality	232 (0.024)
Evading Growth Suppressors	3,362 (0.345)
Avoiding Immune Destruction	752 (0.077)
Deregulating Cellular Energetics	821 (0.084)
avg.	1,749 (0.179)

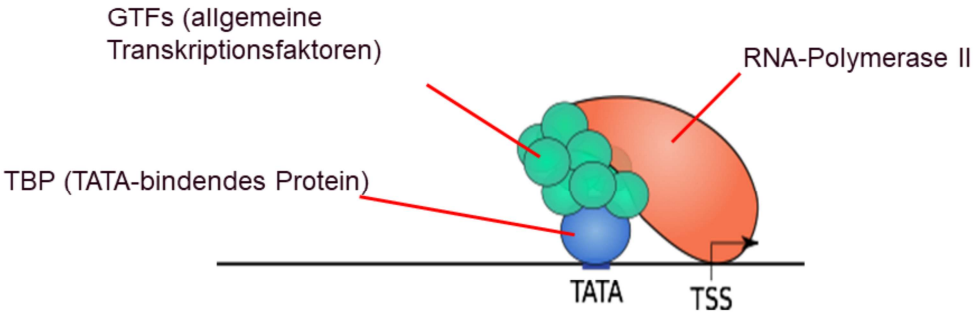
Ein großer Anteil (72%) der anders verschalteten (rewired) Interaktionen betrifft Proteine, die mit den hier aufgeführten, bekannten Merkmalen von Krebs („hallmarks of cancer“) assoziiert sind.

Was für Proteine sind an diesen „verschalteten“ Interaktionen beteiligt? 72% von ihnen, nämlich 7028, sind funktionell mit einem Hallmark-Term für Tumore annotiert. Diese Verschaltungen können daher erwartet werden.

Transkriptionfaktor-komplexe in Hefe und ihre Rolle

In Proteininteraktionsnetzwerken kann man ebenfalls Proteinkomplexe aus mehreren Proteine identifizieren, bzw. zumindest hochwahrscheinliche Kandidaten.

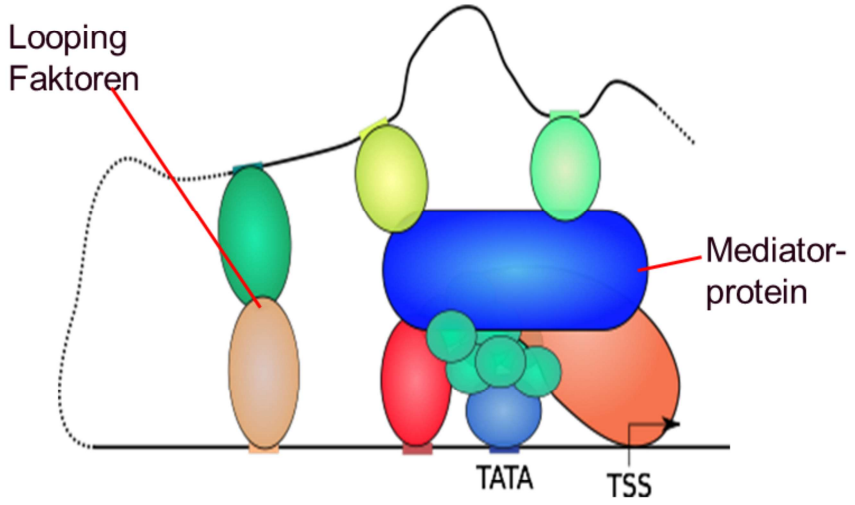
Transkription: Rolle von TFs



TBP bindet an die DNA, wechselwirkt mit den allgemeinen Transkriptionsfaktoren und rekrutiert RNA-Polymerase II

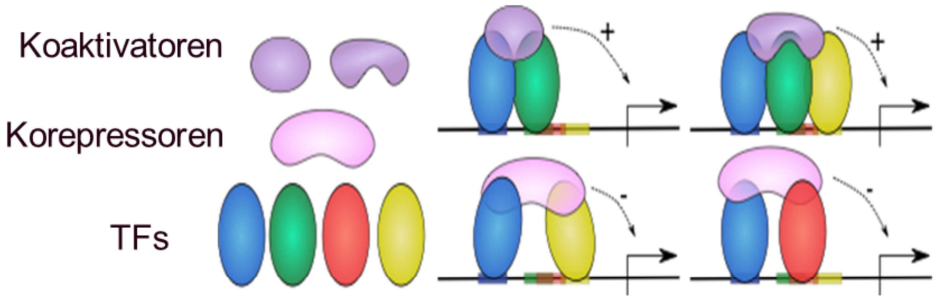
In diesem Beispiel interessierten wir uns für Proteinkomplexe, die mehrere Transkriptionsfaktoren enthalten können und z.B. am Transkriptionsstart von Genen (TSS) binden.

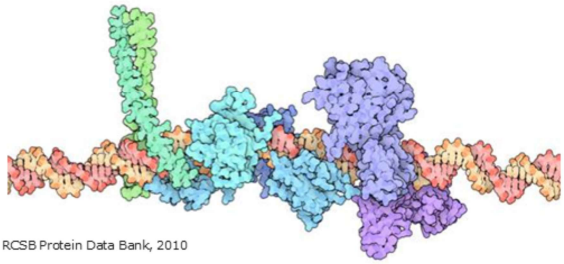
Kombinatorische Vielfalt vieler TFs, Bindung weiterer Proteine möglich



Solche Komplexe können z.B. auch den Mediator-Komplex enthalten und in engem räumlichem Kontakt zu in der genomischen Sequenz weit entfernten Enhancer-Regionen stehen.

cis-regulatorische Module



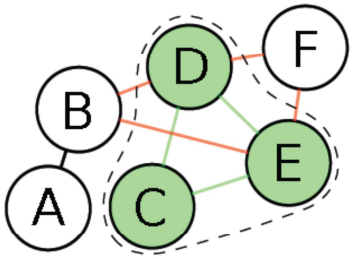


IFN-enhanceosome from RCSB Protein Data Bank, 2010

Das untere Bild zeigt eine Kristallstruktur, in der 10 Transkriptionsfaktoren an die rot/orange DNA gebunden sind.

Im oberen Bild wird veranschaulicht, dass Transkriptionsfaktoren entweder direkt miteinander in Kontakt stehen können (ganz oben), oder über ein drittes Protein in Kontakt stehen können (Mitte).

identifiziere Proteinkomplexe, die TFs beinhalten aus PPI-Netzwerk



Der gestrichelt umrandete Komplex C-D-E hat die maximale Cohesiveness.

BIOINFORMATICS 2014, pages i1–i7 doi:10.1093/bioinformatics/btu448

Identifying transcription factor complexes and their roles

Thorsten Will and Volkhard Helms*

Center for Bioinformatics, Campus Building E2.1, Saarland University, D-66123 Saarbrücken, Germany

Verwende Idee der Methode ClusterOne: Identifiziere Kandidaten für TF-Komplexe im Protein-Interaktionsgraph durch Maximierung der Cohesiveness f(V):

$$f(V) = \frac{w^{in}(V)}{w^{in}(V) + w^{bound}(V)}$$

$w^{in}(V)$: Summe der internen (gewichteten) Kanten, im Beispiel grün

$w^{bound}(V)$: Summe der externen Kanten, im Beispiel orange

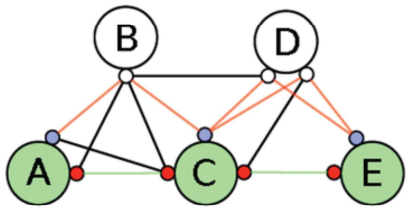
Wir verwendeten als Input ein gewichtetes Proteininteraktionsnetzwerk von PrePPI, bei dem die Gewichte die Konfidenz angeben, ob die Interaktion tatsächlich existiert.

Der Algorithmus ClusterOne identifiziert dann kompakte bzw. kohäsive Regionen, bei denen die Summe der internen Kantengewichte relativ zur Summe aus internen und externen Kantengewichten möglichst groß ist.

In der Abbildung wurde damit der Komplex-Kandidat der 3 Proteine C-D-E identifiziert, die alle miteinander verbunden sind. Das externe Protein F interagiert zwar ebenfalls mit D und E, aber nicht mit C. Deshalb ist f(V) für den Komplex C-D-E-F geringer als für C-D-E.

Domänen-Domänen Repräsentation des PPI-Netzwerks

Annahme: jede Domäne kann nur an einer Interaktion beteiligt sein.



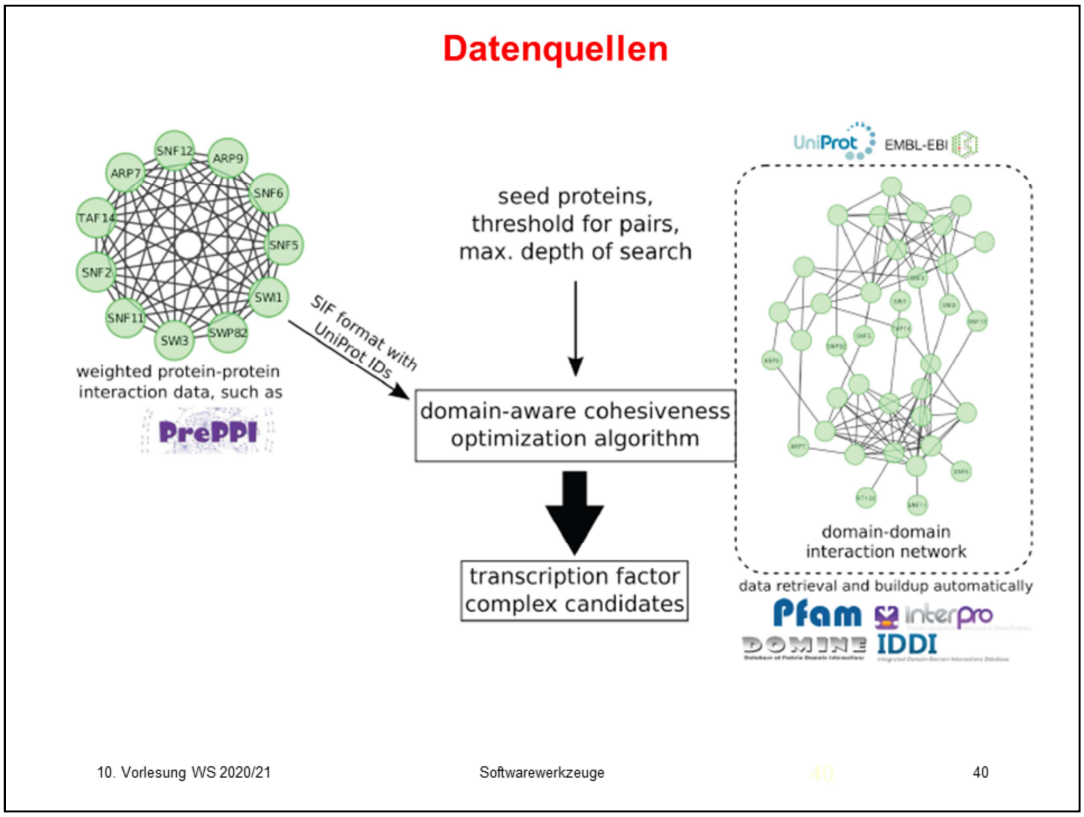
Die grünen Proteine A, C, E bilden aktuellen Komplex.
Die grünen Kanten zwischen ihnen stehen für direkte Kontakte zwischen ihren roten Domänen (kleine Kreise).

B und D sind Kandidaten für einen größeren Komplex.
Ihre weißen Domänen könnten über die orangen Kanten neue Interaktionen mit den nicht belegten (blauen) Domänen von A, C, E ausbilden.

In unserer Methode DACO verwenden wir zudem eine Domänendarstellung der interagierenden Proteine. Die Domänen werden durch kleine Kreise angedeutet.

In diesem Beispiel interagieren die drei Proteine über ihre roten Domänen. Wir nehmen an, dass jede Domäne maximal eine Interaktion ausbilden kann.

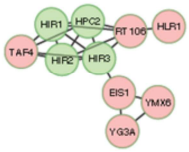
Es gibt weitere Proteine B und D, die mit ihren freien Domänen mit den blauen Domänen der Proteine A, C und E interagieren könnten.



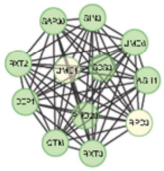
Als Input benötigt unser Algorithmus ein Proteininteraktionsnetzwerk des gewünschten Organismus (links Abbildung) und eine Datenbank für bekannte Wechselwirkungen zwischen Proteindomänen (rechts). Diese Domänen-Kontakte werden aus allen bekannten Organismen kombiniert, da physikalische Interaktionen zwischen Proteindomänen im Allgemeinen sehr gut konserviert sind. Wenn etwa zwei Domänen von zwei menschlichen Proteinen miteinander wechselwirken, gilt dies mit hoher Wahrscheinlichkeit auch für homologe Domänen in einem anderen Organismus. Mit einem systematischen, kombinatorischen Algorithmus erzeugen wir dann für alle bekannten Transkriptionsfaktoren (hier aus Hefe) Komplexe mit anderen Proteinen inklusive weiterer Transkriptionsfaktoren.

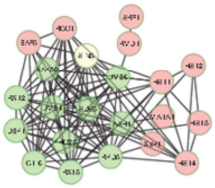
Beispiele für TF-Komplexe (DACA vs. ClusterONE)

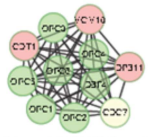
(a) HIR(SGD) / DACO

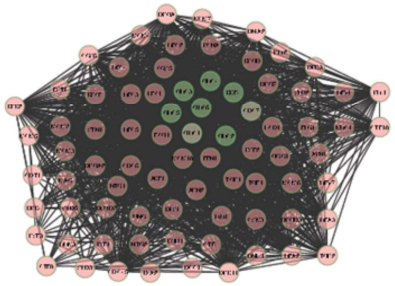


(b) HIR(SGD) / ClusterONE


(c) RPD3L(CYC2008) / DACO


(d) RPD3L(CYC2008) / ClusterONE


(e) ORC(MIPS) / DACO

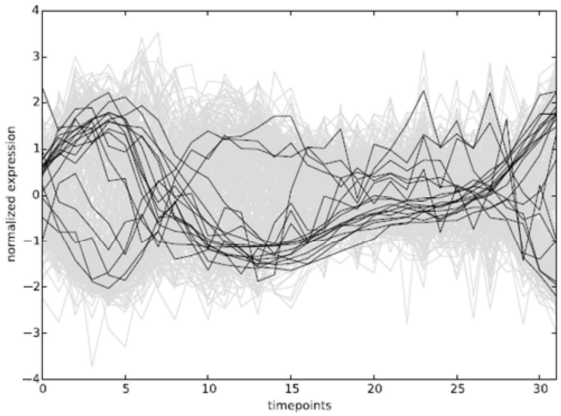

(f) ORC(MIPS) / ClusterONE

Grüne Knoten: Proteine im Referenzkomplex, die korrekt vorhergesagt wurden.

Rote Knoten: weitere Proteine, die als Teil des Komplexes vorhergesagt werden, die aber experimentell nicht bestätigt sind.

Hier vergleichen wir die Lösungen von DACO und des konkurrierenden Algorithmus ClusterONE mit bekannten Proteinkomplexen. Grün gefärbte Knoten sind deren bekannte Komponenten. ClusterONE erzeugt deutlich zu große Komplexe, besonders im Beispiel (f).

Targetgene von TF Komplexen sind ko-exprimiert!



X-Achse: 32 Zeitpunkte entlang des Zellzyklus von Hefezellen.
Zellen wurden für Messung synchronisiert.

Grau: Expression der Targetgene von MET4 **oder** MET32 während des Zellzyklus

Schwarz: Expression der Targetgene, die sowohl von MET4 **und** von MET32 reguliert werden.

-> es gibt 2 Gruppen von Targetgenen, die sich sehr ähnlich verhalten.

In diesem Beispiel sehen wir, welche Einblicke man durch die Berücksichtigung der kombinatorischen Vielfalt von Transkriptionsfaktorkomplexen bekommen kann.

Die Abbildung zeigt die zeitabhängige Variation der Expression von bestimmten Genen in 32 Zeitpunkten entlang des Zellzyklus von Hefe. Hellgrau angedeutet sind alle Gene, die entweder von dem Transkriptionsfaktor MET4 oder von dem TF MET32 reguliert werden. Die Expressionsprofile enthalten kein offensichtliches Muster.

Wenn man aber nur diejenigen Gene plottet (schwarze Linien), die von MET4 und von MET32 reguliert werden (d.h. in ihrem Promoterbereich jeweils eine TFBS-Sequenz (siehe V4) für MET4 und für MET32 enthalten), sieht man zwei Moden von Expressionsprofilen, die entgegengesetzt verlaufen. Diejenigen Gene, die zuerst hochreguliert werden (Maximum bei Zeitpunt 3), sind bei t=12 stark runterreguliert. Diejenigen, die bei t=3 stark runterreguliert sind, sind bei t=12 in einem Maximum.

Zusammenfassung – PP-Komplexe und Netzwerke

Etwa die Hälfte aller zellulären Proteine beteiligen sich transient oder permanent an Interaktionen mit anderen Proteinen.

Im Mittel interagiert ein Protein mit 6 anderen Proteinen (in Hefe).

Protein-Schnittstellen sind (etwas) stärker **konserviert** als die restliche Protein-Oberfläche (Problem: es kann ja weitere Interaktionen geben …).

Korrelierte Mutationen an Schnittstellen sind starke Indizien für PPIs.

Hub-Proteine im Protein-Protein-Interaktionsnetzwerk haben eine höhere Wahrscheinlichkeit **essentiell** zu sein.

Proteinkomplexe, die mehrere Transkriptionsfaktoren erhalten, erhöhen die kombinatorische Vielfalt der Genregulation.

Während der Zelldifferenzierung bzw. Krankheitsentstehung ändern sich eine Vielzahl an Protein-Interaktionen.

In V11 werden wir uns mit metabolischen Netzwerken und evtl. etwas mit Diffusionsprozessen beschäftigen.