V 3 – Data for Building
Protein Interaction Networks

- Detect PPIs by experimental methods
- Detect (predict) PPIs by computational methods
- Derive condition-specific PPIs by data integration

Fri, Nov 4, 2016
Different Roles of Protein Complexes

Assembly of structures
protein machinery is built from parts via dimerization and oligomerization
Complex formation may lead to modification of the active site
Complex formation may lead to increased diversity
Cooperation and allostery
Identification of proteins / components of a complex (1): gel electrophoresis

Electrophoresis: directed diffusion of charged particles in an electric field

- faster diffusion
  - higher charge, smaller
- slower diffusion
  - lower charge, larger

Put proteins in a spot on a gel-like matrix, apply electric field

→ separation according to size (mass) and charge
→ identify constituents of a complex

Nasty details: protein charge vs. pH, cloud of counter ions, protein shape, denaturation, …
**SDS-PAGE**

For better control: denature proteins with detergent

Often used: sodium dodecyl sulfate (**SDS**)  
→ denatures and coats the proteins with a negative charge  
→ charge proportional to mass  
→ traveled distance per time  

\[ x \propto \frac{1}{\log(M)} \]

→ **SDS-polyacrylamide gel electrophoresis**

After the run: **staining** to make proteins visible

For "quantitative" analysis: compare to **marker**  
(set of proteins with known masses)

Image from Wikipedia, marker on the left lane
Protein Charge?

*Protein charge at pH=7*

\[ \approx \sum \text{Lys} + \sum \text{Arg} - \sum \text{Asp} - \sum \text{Glu} + \sum \text{co-factors} \]

Main source for charge differences: pH-dependent protonation states

\[ P = \frac{1}{1 + 10^{\text{pH} - \text{pK}}} \]

pKa = pH value for 50% protonation

Asp 3.7–4.0 … His 6.7–7.1 … Lys 9.3-9.5

Each H\(^+\) has a +1e charge

→ **Isoelectric point**: pH at which the protein is uncharged

→ protonation state cancels permanent charges
2D Gel Electrophoresis

**Two steps:**

i) separation by isoelectric point via pH-gradient

ii) separation by mass with SDS-PAGE

---

**Step 1:**

- Low pH:
  - Protonated => pos. charge
  - Unprotonated => neg. charge

- High pH:

**Step 2:**

SDS-Page

→ Most proteins differ in mass and isoelectric point (pI)
Detect interactions: Yeast Two-Hybrid method

Discover binary protein-protein interactions (bait/prey) via physical interaction

Transcription factor consisting of binding domain (BD) + activator domain (AD) induces expression of reporter gene (LacZ or GFP)

Disrupt BD-AD protein; fuse bait to BD, prey to AD

→ expression only when bait:prey-complex formed

Reporter gene may be fused to green fluorescent protein.
Pros and Cons of Y2H

Advantages:

• *in vivo* test for interactions
• cheap + robust → large scale (genome-wide) tests possible

Problems:

• investigates the interaction between
  (i) overexpressed
  (ii) fusion proteins in the
  (iii) yeast
  (iv) nucleus

• spurious interactions via third protein → many false positives (up to 50% errors)
Identify fragments of proteins / components of a complex (2): Mass Spectrometry

HPLC: high pressure liquid chromatography (first purification step)

Then identify constituents of a (fragmented) complex by MS via their mass/charge patterns m / z

Overview LC-MS

1) Metabolite separation via IC/HPLC

2) Mass detection

3) Extraction of specific masses

http://gene-exp.ipk-gatersleben.de/body_methods.html
Detect interactions:
Tandem affinity purification (also „pull-down“)

Yeast 2-Hybrid-method can only identify binary complexes.

In **affinity purification**, a protein of interest (bait) is tagged with a molecular label (dark route in the middle of the figure) to allow easy purification. The tagged protein is then co-purified together with its interacting partners (W–Z). This strategy can be applied on a genome scale (as Y2H).

Identify proteins by mass spectrometry (MALDI-TOF).

---

TAP analysis of yeast PP complexes

Identify proteins by scanning yeast protein database for protein composed of fragments of suitable mass.

(a) lists the identified proteins according to their localization -> no apparent bias for one compartment, but very few membrane proteins (should be ca. 25%)

(d) lists the number of proteins per complex -> half of all PP complexes have 1-5 members, the other half is larger

(e) Complexes are involved in practically all cellular processes

Validation of TAP methodology

Check of the method: can the same complex be obtained for different choices of the attachment point (tag protein is attached to different components of complex shown in (b))? 

Yes, more or less (see gel in (a)).

< signs mark tag proteins in the gel lane

Pros and Cons of TAP-MS

Advantages:

• **quantitative** determination of complex partners *in vivo* without prior knowledge

• simple method, high yield, high throughput

Difficulties:

• tag may **prevent** binding of the interaction partners

• tag may change (relative) **expression** levels

• tag may be **buried** between interaction partners
  → no binding to beads
Protein interactions in nuclear pore complex

Figure (right) shows 20 NPCs (blue) in a slice of a nucleus. **Aim:** identify individual PPIs in Nuclear Pore Complex.

Below: mutual arrangement of Nup84-complex-associated proteins as visualized by their localization volumes in the final NPC structure. Nup84 protein shown in **light brown.**
**SDS + MS: Composites involving Nup84**

Above lanes: name of ProteinA-tagged protein and identification number for composite

<table>
<thead>
<tr>
<th>Protein</th>
<th>Identification Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nup85</td>
<td></td>
</tr>
<tr>
<td>Nup95</td>
<td></td>
</tr>
<tr>
<td>Nup96</td>
<td></td>
</tr>
<tr>
<td>Nup97</td>
<td></td>
</tr>
<tr>
<td>Nup98</td>
<td></td>
</tr>
<tr>
<td>Nup99</td>
<td></td>
</tr>
<tr>
<td>Nup100</td>
<td></td>
</tr>
<tr>
<td>Nup101</td>
<td></td>
</tr>
<tr>
<td>Nup102</td>
<td></td>
</tr>
<tr>
<td>Nup103</td>
<td></td>
</tr>
<tr>
<td>Nup104</td>
<td></td>
</tr>
<tr>
<td>Nup105</td>
<td></td>
</tr>
<tr>
<td>Nup106</td>
<td></td>
</tr>
<tr>
<td>Nup107</td>
<td></td>
</tr>
<tr>
<td>Nup108</td>
<td></td>
</tr>
<tr>
<td>Nup109</td>
<td></td>
</tr>
<tr>
<td>Nup110</td>
<td></td>
</tr>
<tr>
<td>Nup111</td>
<td></td>
</tr>
<tr>
<td>Nup112</td>
<td></td>
</tr>
<tr>
<td>Nup113</td>
<td></td>
</tr>
<tr>
<td>Nup114</td>
<td></td>
</tr>
<tr>
<td>Nup115</td>
<td></td>
</tr>
<tr>
<td>Nup116</td>
<td></td>
</tr>
<tr>
<td>Nup117</td>
<td></td>
</tr>
<tr>
<td>Nup118</td>
<td></td>
</tr>
<tr>
<td>Nup119</td>
<td></td>
</tr>
<tr>
<td>Nup120</td>
<td></td>
</tr>
<tr>
<td>Nup121</td>
<td></td>
</tr>
<tr>
<td>Nup122</td>
<td></td>
</tr>
<tr>
<td>Nup123</td>
<td></td>
</tr>
<tr>
<td>Nup124</td>
<td></td>
</tr>
<tr>
<td>Nup125</td>
<td></td>
</tr>
<tr>
<td>Nup126</td>
<td></td>
</tr>
<tr>
<td>Nup127</td>
<td></td>
</tr>
<tr>
<td>Nup128</td>
<td></td>
</tr>
<tr>
<td>Nup129</td>
<td></td>
</tr>
<tr>
<td>Nup130</td>
<td></td>
</tr>
<tr>
<td>Nup131</td>
<td></td>
</tr>
<tr>
<td>Nup132</td>
<td></td>
</tr>
<tr>
<td>Nup133</td>
<td></td>
</tr>
<tr>
<td>Nup134</td>
<td></td>
</tr>
<tr>
<td>Nup135</td>
<td></td>
</tr>
<tr>
<td>Nup136</td>
<td></td>
</tr>
<tr>
<td>Nup137</td>
<td></td>
</tr>
<tr>
<td>Nup138</td>
<td></td>
</tr>
<tr>
<td>Nup139</td>
<td></td>
</tr>
<tr>
<td>Nup140</td>
<td></td>
</tr>
<tr>
<td>Nup141</td>
<td></td>
</tr>
<tr>
<td>Nup142</td>
<td></td>
</tr>
<tr>
<td>Nup143</td>
<td></td>
</tr>
<tr>
<td>Nup144</td>
<td></td>
</tr>
<tr>
<td>Nup145</td>
<td></td>
</tr>
<tr>
<td>Nup146</td>
<td></td>
</tr>
<tr>
<td>Nup147</td>
<td></td>
</tr>
<tr>
<td>Nup148</td>
<td></td>
</tr>
<tr>
<td>Nup149</td>
<td></td>
</tr>
<tr>
<td>Nup150</td>
<td></td>
</tr>
<tr>
<td>Nup151</td>
<td></td>
</tr>
<tr>
<td>Nup152</td>
<td></td>
</tr>
<tr>
<td>Nup153</td>
<td></td>
</tr>
<tr>
<td>Nup154</td>
<td></td>
</tr>
<tr>
<td>Nup155</td>
<td></td>
</tr>
<tr>
<td>Nup156</td>
<td></td>
</tr>
<tr>
<td>Nup157</td>
<td></td>
</tr>
</tbody>
</table>

Identity of co-purifying proteins:

**Blue**: PrA-tagged proteins,

**Black**: co-purifying nucleoporins,

**Grey**: NPC-associated proteins,

**Red**: and other proteins (e.g. contaminants)

Affinity-purified PrA-tagged proteins and interacting proteins were resolved by SDS–PAGE and visualized with Coomassie blue. The bands marked by filled circles at the left of the gel lanes were identified by mass spectrometry (cut out band from the gel and use as input for MS).
Indirect Evidence on PPIs: Synthetic Lethality

Apply two mutations that are viable on their own, but lethal when combined.

In cancer therapy, this effect implies that inhibiting one of these genes in a context where the other is defective should be selectively lethal to the tumor cells but not toxic to the normal cells, potentially leading to a large therapeutic window.

<table>
<thead>
<tr>
<th>Gene X</th>
<th>Gene Y</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>No effect</td>
</tr>
<tr>
<td>−</td>
<td>+</td>
<td>No effect</td>
</tr>
<tr>
<td>+</td>
<td>−</td>
<td>No effect</td>
</tr>
<tr>
<td>−</td>
<td>−</td>
<td>Death</td>
</tr>
</tbody>
</table>

Synthetic lethality may point either to:
• physical interaction of proteins (they are building blocks of a complex)
• both proteins belong to the same pathway
• both proteins have the same function (redundancy)
Indirect Evidence on PPIs: Gene Coexpression

All constituents of a complex should be present at the same point in the cell cycle → test for correlated expression

Co-expression is not a direct indication for formation of complexes (there are too many co-regulated genes), but it is a useful "filter"-criterion.

Standard tools: DNA micro arrays / RNA-seq

DeRisi, Iyer, Brown, Science 278 (1997) 680:

Diauxic shift from fermentation (growth on sugar) to respiration (growth on ethanol) in S. cerevisiae

→ Identify groups of genes with similar expression profiles
# Interaction Databases

Bioinformatics: make experimental data available in databases

### Table 3.1: Some public databases compiling data related to protein interactions. (P) and (D) stand for proteins and domains (the number of interactions reflects the status of June 2007).

<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
<th>Number of interactions</th>
<th>Type</th>
<th>Proteins/dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIPS</td>
<td>mips.gsf.de/genome/prot/impact</td>
<td>4300</td>
<td>curated</td>
<td>(P)</td>
</tr>
<tr>
<td>BIND</td>
<td>bind.unleashedinformatics.com</td>
<td>200000</td>
<td>curated</td>
<td>(P)</td>
</tr>
<tr>
<td>MINT</td>
<td>160.80.34.4/mint/</td>
<td>103800</td>
<td>curated</td>
<td>(P)</td>
</tr>
<tr>
<td>DIP</td>
<td>dip.doe-mbi.ucla.edu</td>
<td>56000</td>
<td>curated</td>
<td>(P)</td>
</tr>
<tr>
<td>PDB</td>
<td><a href="http://www.rcsb.org/pdb">www.rcsb.org/pdb</a></td>
<td>800 complexes</td>
<td>curated</td>
<td>(P, D)</td>
</tr>
<tr>
<td>HPRD</td>
<td><a href="http://www.hprd.org">www.hprd.org</a></td>
<td>37500</td>
<td>automated</td>
<td>D</td>
</tr>
<tr>
<td>ScoPpi</td>
<td><a href="http://www.scoppi.org">www.scoppi.org</a></td>
<td>102000</td>
<td>integrated</td>
<td>(P)</td>
</tr>
<tr>
<td>UniHII</td>
<td>theoederich.fz.mdcb-berlin.de:8080/unihi/home</td>
<td>209000</td>
<td>integrated data</td>
<td>(P)</td>
</tr>
<tr>
<td>STRING</td>
<td>string.embl.de</td>
<td>interactions of 1500000 proteins</td>
<td>integrated data from genomic context, high-throughput experiments, coexpression, previous knowledge data extracted from PDB</td>
<td>(P)</td>
</tr>
</tbody>
</table>
Initially low overlap of results

For **yeast**: ~6000 proteins  =>  ~18 million potential interactions
rough estimates:  \( \leq 100000 \) interactions occur

→ 1 true positive for 200 potential candidates  =  **0.5%**
→ **decisive** experiment must have **accuracy** \(< < 0.5\%\) false positives

**Different experiments** detect different interactions:
For yeast: 80000 interactions known in 2002
only 2400 were found by \( \geq 2 \) experiments

Problems with experiments:

i) incomplete coverage

ii) (many) false positives

iii) selective to type of interaction
    and/or compartment

Y2H: yeast two hybrid screen
TAP: tandem affinity purification
HMS-PCI: protein complex identification by MS

von Mering (2002)
Criteria for reliability of detected PPIs

Guiding principles to judge experimental results on PPIs (incomplete list!):

1) check **mRNA abundance** of detected PPIs:
   most experimental techniques are biased towards high-abundance proteins.
   If this is the case, results for low-abundance proteins are not reliable.

2) Check localization to cellular **compartments**:
   • most methods have their "preferred compartment"
   • if interacting proteins belong to the same compartment
     => results are more reliable

3) **co-functionality**
   it is realistic to assume that members of a protein complex should have closely
   related biological functions -> check whether interaction proteins have
   overlapping annotations with terms from Genome Ontology (GO)
In-Silico Prediction Methods

**Sequence**-based:
- gene clustering
- gene neighborhood
- Rosetta stone
- phylogenetic profiling
- coevolution

**Structure**-based:
- interface propensities
- protein-protein docking
- spatial simulations (e.g. MD)

"Work on the parts list"
→ fast
→ unspecific
→ high-throughput methods for pre-sorting

"Work on the parts"
→ specific, detailed
→ expensive
→ accurate

Will be covered today
Not subject of this lecture
Gene Clustering

**Idea:** functionally related proteins or parts of a complex are expressed **simultaneously**

Search for genes with a **common promoter** → when activated, all are transcribed together as one **operon**

**Example:**
bioluminescence in *V. fischeri* is regulated via quorum sensing → three proteins: I, AB, CDE are responsible for this. They are organized as 1 operon named *luxICDABE.*
**Gene Neighborhood**

**Hypothesis** again: functionally **related** genes are expressed **together**

"functionally related" means same \{complex | pathway | function | …\}

→ Search for **similar arrangement** of related genes in **different organisms**

(<=) Gene clustering: done in one species, need to know promoters
**Rosetta Stone Method**

**Idea:** find homologous genes ("**words**") in genomes of different organisms ("**texts**")
- check if **fused gene pair** exists in one organism
  → May indicate that these 2 proteins form a complex

Multi-lingual stele from 196 BC, found by the French in 1799
The same decree is inscribed on the stone 3 times, in hieroglyphic, demotic, and greek.
→ key to deciphering meaning of hieroglyphs

Enright, Ouzounis (2001):
40000 predicted pair-wise interactions from search across 23 species
**Phylogenetic Profiling**

**Idea**: either all or none of the proteins of a complex should be present in an organism

→ compare presence of protein homologs across species (e.g., via sequence alignment)

![Diagram of protein homologs across species](image)
Distances in Phylogenetic Profiling

Hamming distance between species: number of different protein occurrences

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>P3</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>P4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P6</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Two pairs with similar occurrence: P2-P7 and P3-P6
These are candidates to interact with each other.
Co-evolution

Binding interfaces of complexes are often better conserved in evolution than the rest of the protein surfaces.

Idea of Pazos & Valencia (1997): if a mutation occurs at one interface that changes the character of this residue (e.g. polar $\rightarrow$ hydrophobic), a corresponding mutation could occur at the other interface at one of the residues that is in contact with the first residue.

Detecting such correlated mutations could help in identifying binding candidates.
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

<table>
<thead>
<tr>
<th></th>
<th>our method</th>
<th>metaPPI</th>
<th>meta-PPISP</th>
<th>PPI-Pred</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td>Acc</td>
<td>Cov</td>
<td>Acc</td>
<td>Cov</td>
</tr>
<tr>
<td>E-1</td>
<td>73%</td>
<td>65%</td>
<td>61%</td>
<td>37%</td>
</tr>
<tr>
<td>others</td>
<td>55%</td>
<td>57%</td>
<td>41%</td>
<td>22%</td>
</tr>
<tr>
<td>overall</td>
<td>63%</td>
<td>61%</td>
<td>49%</td>
<td>28%</td>
</tr>
</tbody>
</table>

*E-1 is type of enzyme–inhibitor.
Toward condition-specific protein interaction networks

Full interaction PP network, e.g. of human = collection of pairwise interactions compiled from different experiments

broad range of applications
But protein interactions can be …

dynamic in time and space

from Han et al., Nature, 2004

same color = similar expression profiles

interaction data itself generally static

condition-specific protein composition

Human tissues from www.pharmaworld.pk
Alzheimer from www.alz.org
Simple condition-specific PPI networks

database(s)  ...  complete protein interaction network

idea: prune to subset of expressed genes

e.g.:
Bossi and Lehner, Mol. Syst. Bio., 2009
Lopes et al., Bioinformatics, 2011
Barshir et al., PLoS CB, 2014
Differential PPI wiring analysis

112 matched normal tissues (TCGA)  112 breast cancer tissues (TCGA)

comparison 1:

\[ P_1 \rightarrow P_2 \rightarrow P_3 \]
\[ P_4 \rightarrow P_5 \]

\[ d_1 \]

comparison 2:

\[ P_1 \rightarrow P_2 \rightarrow P_3 \]
\[ P_4 \rightarrow P_5 \]

\[ d_2 \]

comparison 3:

\[ P_1 \rightarrow P_2 \rightarrow P_3 \]
\[ P_4 \rightarrow P_5 \]

\[ d_3 \]

\[ \sum d_i \]

Check whether rewiring of a particular PP interaction occurs in a significantly large number of patients compared to what is expected by chance rewiring events.

Will, Helms, Bioinformatics, 47, 219 (2015)
doi: 10.1093/bioinformatics/btv620
Coverage of PPIs with domain information

Standard deviations reflect differences between patients.

About 10,000 out of 133,000 protein-protein interactions are significantly rewired between normal and cancer samples.

<table>
<thead>
<tr>
<th>GENE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>avg. number of proteins (normal)</td>
<td>12,678 ± 223</td>
</tr>
<tr>
<td>avg. number of proteins (tumor)</td>
<td>12,528 ± 206</td>
</tr>
<tr>
<td>avg. number of interactions (normal)</td>
<td>134,348 ± 2,387</td>
</tr>
<tr>
<td>avg. number of interactions (tumor)</td>
<td>133,128 ± 2,144</td>
</tr>
<tr>
<td>( P_{rew} )</td>
<td>0.067 ± 0.016</td>
</tr>
<tr>
<td>significantly rewired interactions</td>
<td>9,754</td>
</tr>
</tbody>
</table>

Table S7: Results obtained using the BioGRID interaction data and using either gene- or various transcript-based network construction approaches. The given numbers denote the sizes of the constructed networks. For all deterministic approaches the standard deviation across all 112 matched samples is shown, for the randomized approach the deviation shown is the average of standard deviations per run. A part of the results for \( P_{rew} \) and significantly rewired interactions are also shown in the upper half of Table 3 in the main text. Both net loss of proteins and interactions from normal to tumor were significant according to a two-sided Wilcoxon signed-rank test applied to the matched pairs of samples. For the
Rewired PPIs are associated with hallmarks

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

Table S10: Results for the rewiring analysis of the BioGRID network in terms of rewired interactions that affect proteins associated with hallmarks of cancer as defined by [1]. A protein interaction was considered relevant regarding a hallmark term if at least one of its associated proteins was part of the corresponding set of hallmark proteins. The results for individual hallmark terms are reported as the absolute quantity of matches (left number) and as fraction of the total number of rewired interactions listed in the first row (in brackets).

A large fraction (72%) of the rewired interactions affects genes that are associated with „hallmark of cancer“ terms.
Not considered yet: alternative splicing

DNA

exon 1  exon 2  exon 3  exon 4

transcription

primary RNA transcript

5'  3'

3'  5'

alternative splicing (~95% of human multi-exon genes)

mRNAs

translation

protein isoforms

AS affects ability of proteins to interact with other proteins
PPIXpress uses domain information

see http://sourceforge.net/projects/ppixpress

I. Determine “building blocks“ for all proteins

- transcript abundance from RNA-seq data
- protein domain composition from sequence (Pfam annotation)

II. Connect them on the domain-level

Use info from high-confidence domain-domain interactions

protein-protein interaction network

domain-domain interaction network

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

mapping:
protein-protein interaction
establish one-to-at-least-one relationship
domain-domain interaction

reference: principal protein isoforms = longest coding transcript
PPIXpress method

I. mapping

II. instantiation

reference: principal protein isoforms

built using most abundant protein isoforms
Differential PPI wiring analysis at domain level

112 matched normal tissues (TCGA) 112 breast cancer tissues (TCGA)

comparison 1:

comparison 2:

comparison 3:

\[ \sum d_i \]

one-tailed binomial test + BH/FDR (<0.05)
Coverage of PPIs with domain information

<table>
<thead>
<tr>
<th>protein set</th>
<th>size of set</th>
<th>matched PPIs</th>
<th>contributing proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete network*</td>
<td>15086</td>
<td>0.264</td>
<td>0.517</td>
</tr>
<tr>
<td>all HM</td>
<td>4407</td>
<td>0.280</td>
<td>0.684</td>
</tr>
<tr>
<td>non HM</td>
<td>10679</td>
<td>0.227</td>
<td>0.449</td>
</tr>
</tbody>
</table>

Domain information is currently available for 51.7% of the proteins of the PP interaction network.

This means that domain information supports about one quarter (26.7%) of all PPIs.

All other PPIs were connected by us via artificially added domains (1 protein = 1 domain).
At domain-level, slightly more (10.111 vs. 9.754) PPIs out of 133,000 PPIs are significantly rewired between normal and cancer samples.

Table S7: Results obtained using the BioGRID interaction data and using either gene- or various transcript-based network construction approaches. The given numbers denote the sizes of the constructed networks. For all deterministic approaches the standard deviation across all 112 matched samples is shown, for the randomized approach the deviation shown is the average of standard deviations per run. A part of the results for $P_{rew}$ and significantly rewired interactions are also shown in the upper half of Table 3 in the main text.

Both net loss of proteins and interactions from normal to tumor were significant according to a two-sided Wilcoxon signed-rank test applied to the matched pairs of samples. For the ALL-DDI construction in BioGRID, for example, there were less proteins in the tumor PPINs with $p < 5.9 \times 10^{-8}$ (GENE: $p < 3.6 \times 10^{-8}$) and less interactions with $p < 3.8 \times 10^{-6}$ (GENE: $p < 3.9 \times 10^{-6}$), respectively.
Rewired PPIs are associated with hallmarks

<table>
<thead>
<tr>
<th></th>
<th>GENE</th>
<th>ALL-DDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>rewired interactions</td>
<td>9,754</td>
<td>10,111</td>
</tr>
<tr>
<td>participation in any hallmark term</td>
<td>7,028</td>
<td>7,343</td>
</tr>
<tr>
<td>fraction in any hallmark term</td>
<td>0.721</td>
<td>0.726</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Resisting Cell Death</td>
<td>4,064 (0.417)</td>
<td>4,316 (0.427)</td>
</tr>
<tr>
<td>Activating Invasion and Metastasis</td>
<td>2,244 (0.230)</td>
<td>2,285 (0.226)</td>
</tr>
<tr>
<td>Sustaining Proliferative Signaling</td>
<td>3,964 (0.406)</td>
<td>4,142 (0.410)</td>
</tr>
<tr>
<td>Inducing Angiogenesis</td>
<td>169 (0.017)</td>
<td>172 (0.017)</td>
</tr>
<tr>
<td>Tumor-Promoting Inflammation</td>
<td>516 (0.053)</td>
<td>537 (0.053)</td>
</tr>
<tr>
<td>Genome Instability and Mutation</td>
<td>1,362 (0.140)</td>
<td>1,419 (0.140)</td>
</tr>
<tr>
<td>Enabling Replicative Immortality</td>
<td>232 (0.024)</td>
<td>360 (0.036)</td>
</tr>
<tr>
<td>Evading Growth Suppressors</td>
<td>3,382 (0.345)</td>
<td>3,557 (0.352)</td>
</tr>
<tr>
<td>Avoiding Immune Destruction</td>
<td>752 (0.077)</td>
<td>772 (0.076)</td>
</tr>
<tr>
<td>Deregulating Cellular Energetics</td>
<td>821 (0.084)</td>
<td>850 (0.084)</td>
</tr>
<tr>
<td><strong>avg.</strong></td>
<td>1,749 (0.179)</td>
<td>1,841 (0.182)</td>
</tr>
</tbody>
</table>

Table S10: Results for the rewiring analysis of the BioGRID network in terms of rewired interactions that affect proteins associated with hallmarks of cancer as defined by [1]. A protein interaction was considered relevant regarding a hallmark term if at least one of its associated proteins was part of the corresponding set of hallmark proteins. The results for individual hallmark terms are reported as the absolute quantity of matches (left number) and as fraction of the total number of rewired interactions listed in the first row (in brackets).

The construction at transcript-level also found a larger fraction (72.6 vs 72.1%) of differential interactions that can be associated with hallmark terms than the gene-level based approach.
The enriched terms that are exclusively found by the transcript-level method (right) are closely linked to carcinogenetic processes.

Hardly any significant terms are exclusively found at the gene level (left).
Conclusion (PPIXpress)

About 10,000 out of 130,000 PP interactions are rewired in cancer tissue compared to matched normal tissue due to altered gene expression.

The method PPIXpress exploits domain interaction data to adapt protein interaction networks to specific cellular conditions at transcript-level detail.

For the example of protein interactions in breast cancer this increase in granularity positively affected the performance of the network construction compared to a method that only makes use of gene expression data.

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

What you learned **today**: how to get some data on PP interactions

- SDS-PAGE
- TAP
- MS
- Y2H
- micro array
- synthetic lethality
- Rosetta stone
- gene clustering
- gene neighborhood
- phylogenetic profiling
- coevolution

**Next lecture**: Mon, Nov. 7, 2016

- combining weak indicators: Bayesian analysis
- identifying communities in networks