Computational methods for predicting protein-protein interactions

Tomi Peltola

T-61.6070 Special course in bioinformatics I

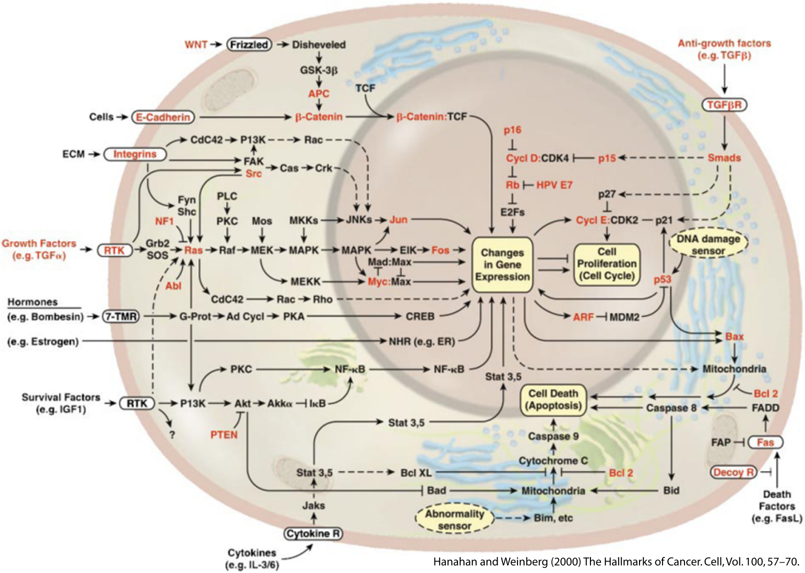
3.4.2008

Outline

- Biological background
    - Protein-protein interactions
- Computational methods
- A model for prediction of protein-protein interactions from sequence alignments
- Summary

Biological background

Biological background – Cell



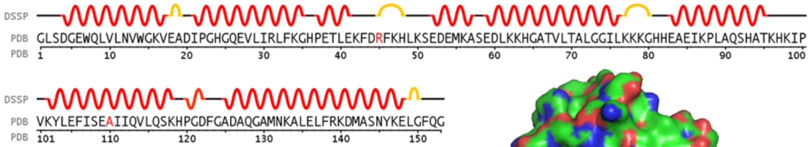
Hanahan and Weinberg (2000) The Hallmarks of Cancer. Cell, Vol. 100, 57–70.

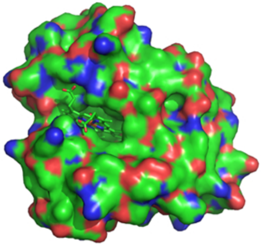
Biological background – Proteins

- Proteins determine the outcome of most cellular processes.
- Cellular functions:
  - enzymes
  - structural and mechanical elements
  - signalling and transport

Biological background – Proteins

- Linear sequences of 20 (standard) amino acids (primary structure).
- Fold into 3D shapes. Shape is important for function.


DSSP
PDB GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIP
PDB 1 10 20 30 40 50 60 70 80 90 100
DSSP
PDB VKYLEFISEAIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
PDB 101 110 120 130 140 150



Myoglobin, PDB ID: 2MM1

Protein-protein interactions

- Possibilities include:
  - Predicting functions of proteins.
  - Predicting protein complexes.
  - Pathways for basic understanding and drug development.
  - Network structure analysis.
- Things to consider:
  - Functional interaction vs. physical interaction.
  - Time scale: transient interactions vs. complexes.
  - Network scale: genome wide, functional modules or pathways.

Protein-protein interactions

- Experimental methods (high-throughput):
  - Yeast two-hybrid.
  - Affinity purification-MS.
  - DNA and protein microarrays.
  - Synthetic lethality.
  - Phage display.
- Databases
  - There's numerous...
  - The International Molecular Exchange Consortium (IMEx).

Computational methods

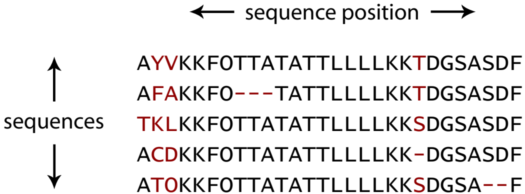
Methods – Basic concepts

- Homology: a relationship of common descent between any entities (in particular genes).
- Orthologs: genes derived from a single ancestral gene in the last common ancestor of the compared species.
- Paralogs: genes related via duplication.

(Koonin (2005) Annual Review of Genetics. Vol. 39: 309–338.)

Methods – Basic concepts

- Sequence alignment: comparing two or more sequences by searching for character patterns that are in the same order in the sequences.



sequence position

sequences

```
AYVKKFOTTATATTLLLLKKTDGSASDF
AFAKKFO---TATTLLLLKKTDGSASDF
TKLKKFOTTATATTLLLLKKSDGSASDF
ACDKKFOTTATATTLLLLKK-DGSASDF
ATOKKFOTTATATTLLLLKKSDGSA--F
```

...and then the actual methods

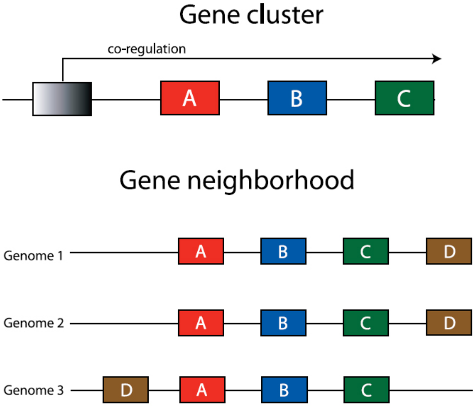
Shoemaker and Panchenko
(2007)
Deciphering Protein-Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners

Figures in this section are from the article

(Rosetta Stone figure is an adapted version).

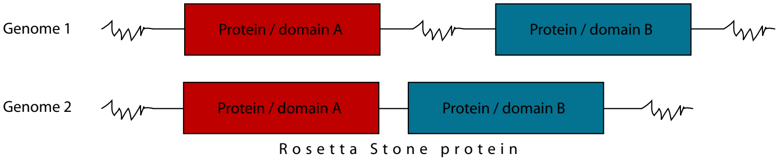
Methods – Genomic distances

- Gene neighbor and gene cluster methods:
  - Operons in bacteria.
  - Co-regulation in eukaryotes.
- Prediction based on intergenic distances.



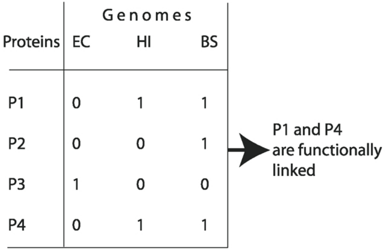
Methods – Rosetta Stone

- Interacting proteins can have fused homologs in other genomes.



Methods – Phylogenetic profile

- Hypothesis: functionally linked or interacting nonhomologous proteins co-evolve
- and have orthologs in other organisms.

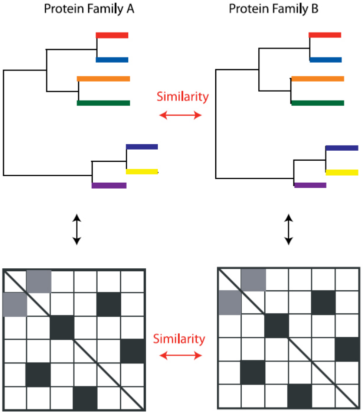


	Genomes		
Proteins	EC	HI	BS
P1	0	1	1
P2	0	0	1
P3	1	0	0
P4	0	1	1

P1 and P4 are functionally linked

- Fully sequenced genomes needed.

Methods – Sequence co-evolution



- Correlated changes in co-evolving proteins.
- ”Phylogenetic substraction” to account for background similarity.

Methods – Classification methods

- Any classification method could be applied:
  - Random Forest Decision
  - Support Vector Machines
  - ...
- Training set needed.
- Feature data: domains, experimental data etc.
- Can easily integrate multiple data sources.

Methods – Problems

- Poor coverage.
- Poor overlap between methods.
- Hard to distinguish between physical and functional relationship.
- Hard to validate
  - Lack of accurate data sets for validation.
  - Methods might not provide confidence estimation.

A model for prediction of protein-protein interactions from sequence alignments

Burger and van Nimwegen
(2008)
Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method

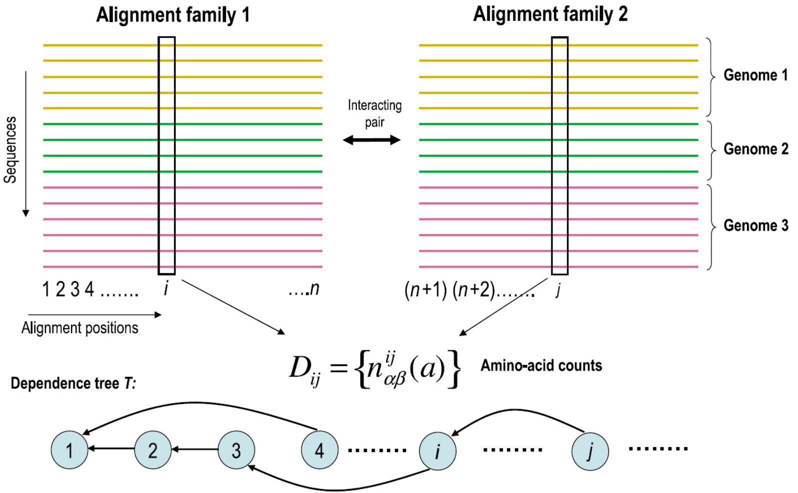
Figures in this section are from the article

(except the one on the computation slide).

BN Model

- Infers interaction partners using multiple sequence alignments of protein families that are known to interact.
- Based on the assumption of co-varying residue pairs for interacting proteins:
    - The identity of a residue is dependent on the identity of one other residue.
    - All possible dependencies are summed over.
- No training set needed. No tunable parameters.

BN Model

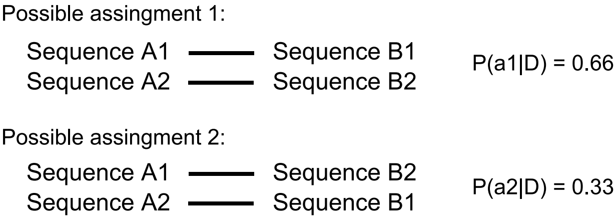


$$D_{ij} = \left\{ n^{ij}_{\alpha\beta}(a) \right\}$$

$$P(D|a,T) = P(D_1)P(D_2|D_1)P(D_3|D_2)P(D_4|D_1)\cdots P(D_i|D_3)\cdots P(D_j|D_i)\cdots$$

BN Model – Computation

- Probability of assignment: $P(a|D)$.
- If we had two sequences per protein family:



Possible assignment 1:

Sequence A1 — Sequence B1

Sequence A2 — Sequence B2

P(a1|D) = 0.66

Possible assignment 2:

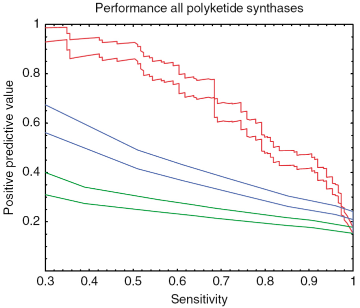
Sequence A1 — Sequence B2

Sequence A2 — Sequence B1

P(a2|D) = 0.33

- If there are 20 sequences per family, there are some 2.4×10^{18} different possible assignments.

BN Model – Results



$$PPV = \frac{true\ positives}{true\ positives + false\ positives}$$

$$sensitivity = \frac{true\ positives}{true\ positives + false\ negatives}$$

BN Model – Summary

- Computationally complex:
  - Gibbs sampling.
  - If summing over dependency trees is intractable, ML-estimated tree can be used.
  - Training set can be included by fixing those assingments.
- Can be extended for several protein families in parallel and unassigned members.
- No tunable parameters:
  - Predictions depend on informative positions in the alignments.
- Generally applicable to multiple sequence alignments.

Summary

- Protein-protein interactions are essential in cellular processes.
- Consideration is needed to what is meant by interaction.
- Computational and experimental methods complement each other.
  - Currently both are limited in applicability and performance.
  - Many possible methods based on different biological principles.
- Analyzing the protein-protein interaction results might be a demanding task in itself.

References

- Shoemaker and Panchenko (2007) *Deciphering Protein-Protein Interactions. Part I. Experimental techniques and databases.* PLoS Comp Biol 3: e42.
- Shoemaker and Panchenko (2007) *Deciphering Protein-Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners.* PLoS Comp Biol 3: e43.
- Burger and van Nimwegen (2008) *Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method.* Molecular Systems Biology 4:165.