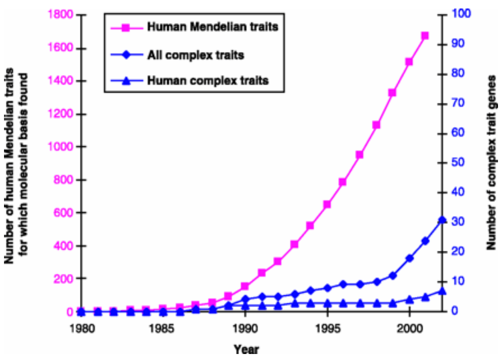
Computational methods for prediction of protein-protein interactions and disease genes

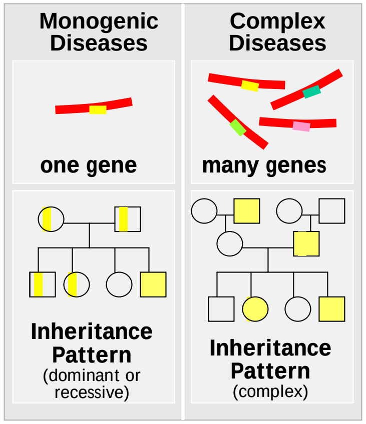
Tu-Bao Ho and Thanh-Phuong Nguyen

Japan Advanced Institute of Science and Technology

From genes to phenotype



Complex diseases: the study remains challenging (association studies).



Monogenic diseases: Correlation between mutations in the patient genome and the symptoms might not be clear (linkage analysis).

Outline

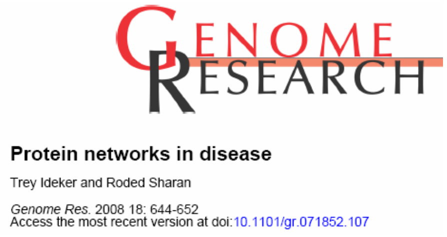
Protein networks in disease

Prediction of protein-protein interactions

PPI-based prediction of disease genes

Protein networks in disease

- Shifted from understanding networks encoded by model species to understanding the networks underlying human disease.


GENOME RESEARCH

Protein networks in disease

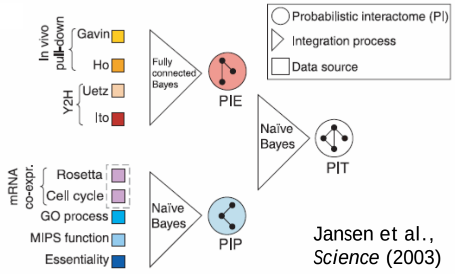
Trey Ideker and Roded Sharan

Genome Res. 2008 18: 644-652
Access the most recent version at doi:10.1101/gr.071852.107

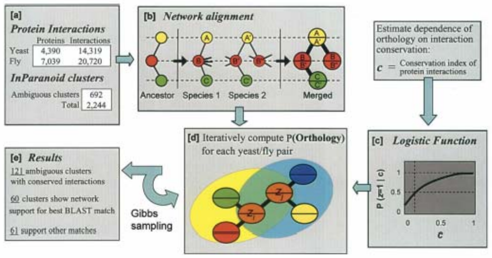
- Four major areas of protein network in disease:
  - ➔ The study of network properties
  - ➔ Identifying new disease genes
  - ➔ Identifying disease-related subnetworks
  - ➔ Network-based disease classification

Network analysis in yeast: a brief tour

- From raw interaction measurements to **higher confidence networks** with quantitative measures.
- **Predict new annotations for proteins**, such as protein function, localization, and functional orthology, etc.
- A third set of methods:
  - ➔ **Synthesize** global properties of biology by analyzing interaction networks.
  - ➔ **Decompose** or partition networks into smaller building blocks



Jansen et al., *Science* (2003)

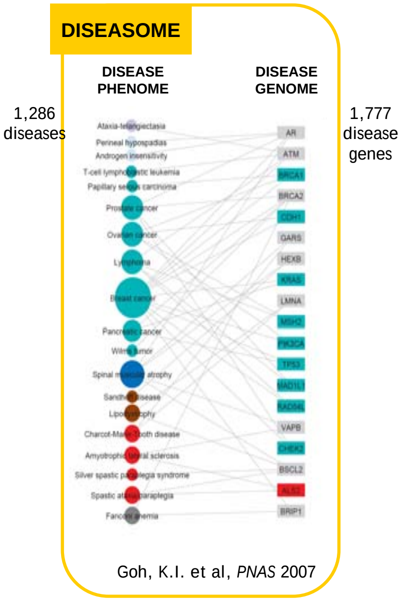


Bandyopadhyay et al., *Genome Res* (2006)

Human network analysis: disease genes properties

Inspired by the findings for yeast, several groups focus on **phenotypes related to human disease**.

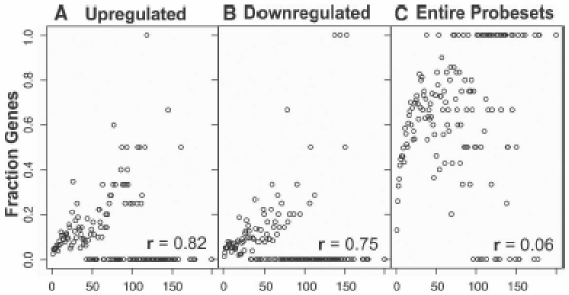
- Jonsson and Bates (2006): 346 human cancer gene network: have twice as many interaction partners as non-cancer proteins.
- Goh (2007): human disease & human gene association network, each genetic disease is connected to the genes known to cause it.



Goh, K.I. et al, *PNAS* 2007

Human network analysis: Overriding conclusion on disease genes properties

Wachi et al., Bioinformatics (2005)



Genes associated with a particular phenotype or function, including the progression of disease, are **not randomly positioned in the network**.

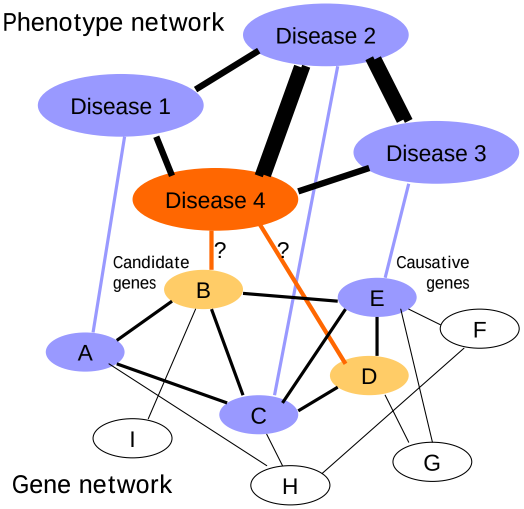
Rather, they tend to exhibit **high connectivity**, **cluster together**, and **occur in central network locations.**

Prediction of disease-causing genes

Key assumption

A network-neighbor of a disease-causing gene is likely to cause either the same or a similar disease

(Goh et al. 2007;
Oti and Brunner 2007).



(Reproduced from www.blackwell-synergycom)

Prediction of disease-causing genes

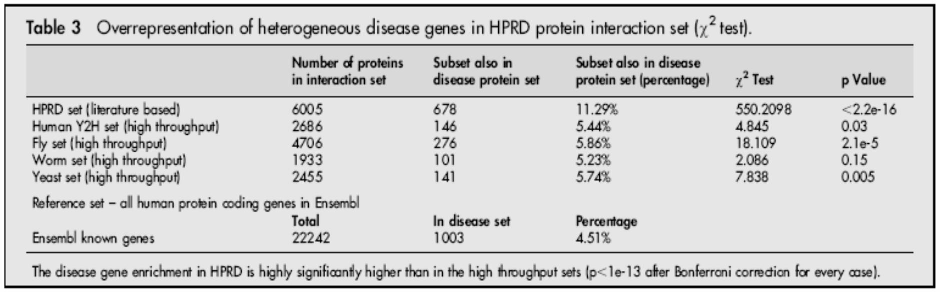


Table 3 Overrepresentation of heterogeneous disease genes in HPRD protein interaction set (χ^2 test).

	Number of proteins in interaction set	Subset also in disease protein set	Subset also in disease protein set (percentage)	χ^2 Test	p Value
HPRD set (literature based)	6005	678	11.29%	550.2098	<2.2e-16
Human Y2H set (high throughput)	2686	146	5.44%	4.845	0.03
Fly set (high throughput)	4706	276	5.86%	18.109	2.1e-5
Worm set (high throughput)	1933	101	5.23%	2.086	0.15
Yeast set (high throughput)	2455	141	5.74%	7.838	0.005
Reference set – all human protein coding genes in Ensembl					
	Total	In disease set	Percentage		
Ensembl known genes	22242	1003	4.51%		

The disease gene enrichment in HPRD is highly significantly higher than in the high throughput sets (p<1e-13 after Bonferroni correction for every case).

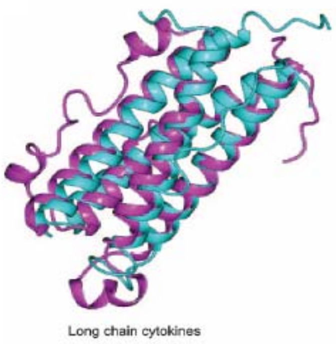
Oti et al., Med. Genet. (2006)

- Oti et al. (2006): those that fell at significant loci and had a protein interaction with a gene already well known to cause disease.
- Lage et al. (2007): phenotype similarity score and used it to look for protein complexes whose genes were associated with similar phenotypes.

Prediction of disease-causing genes

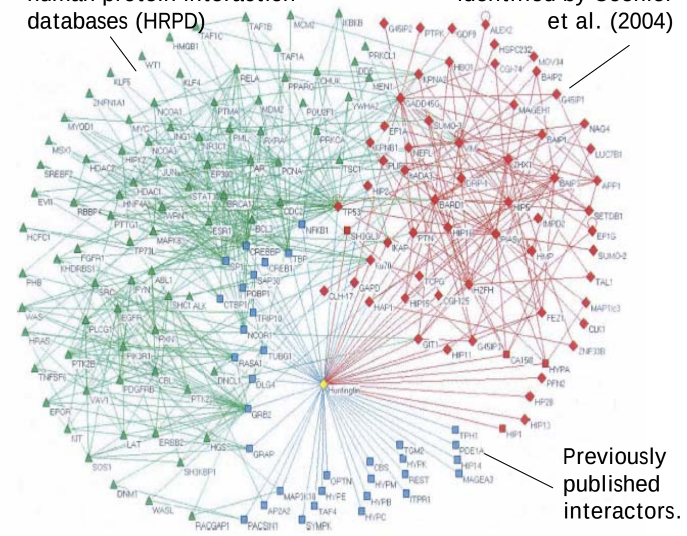
The idea that proteins close to one another in a network cause similar diseases is becoming an increasingly important factor in the hunt for disease genes.

- All approaches involve **superimposing a set of candidate genes alongside a set of known disease genes on a physical or functional network**.
- "De-novo" approaches that do not depend on prior knowledge of disease genes are yet to be developed.



Disease-related subnetworks identification

- Concrete hypotheses as to the molecular complexes, signaling pathways, etc.
- Goeher (2004): PPI subnetworks around HTT, mutations that cause Huntington disease



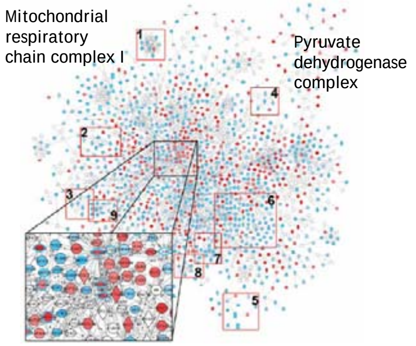
Interactors culled from human protein interaction databases (HRPD)

Y2H interactors newly identified by Goehler et al. (2004)

Previously published interactors.

Disease-related subnetworks identification

- Overlaying expression profiles as states on a functional network (Calvano, 2005).
- Proteins are linked based on coexpression, phenotypic similarity, and genetic or physical interactions (Pujana et al. 2007).


Mitochondrial respiratory chain complex I
Pyruvate dehydrogenase complex
1
2
3
4
5
6
7
8
9

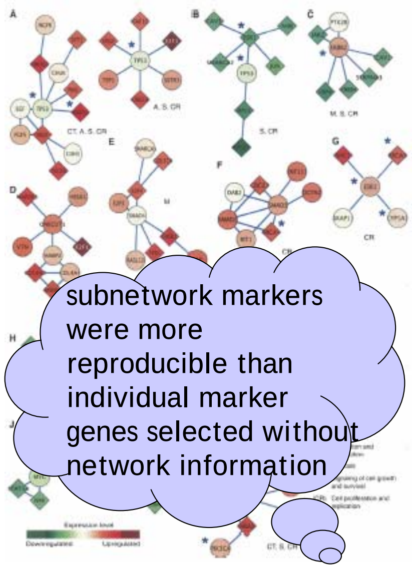
Calvano et al., Nature (2005)

Integrating disease genes with physical or functional networks can lead to the identification of additional disease-related genes and generate subnetworks that offer mechanistic hypotheses about the causes of disease.

Network-based classification of case-control studies

Biomarker identification by case-control classification: Quackenbush (2006), Sotiriou and Piccart (2007), Chuang et al. (2007), etc.

Typically, one superimposes gene-expression data onto the network to identify links, or more composite subnetwork structures, whose aggregate expression discriminates between disease states.

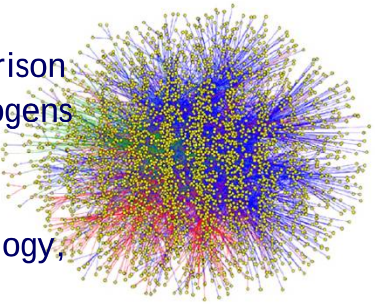


subnetwork markers were more reproducible than individual marker genes selected without network information

Chuang et al., *Mol. Syst. Biol.* (2007)

The future of networks and disease

Typical roads ahead:



- Protein network evolutionary comparison
- Network-level analyses of viral pathogens
- Effects of genetic and environmental perturbations on human populations
- Network-based analysis in pharmacology, i.e., drug discovery and targeting

The recent availability of human molecular interaction networks has revolutionized studies on single genes by demonstrating the importance not only of the proteins themselves, but of their inter-relationships.

Databases with disease annotation

- **OMIM** (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM)
  - ➔ a catalog of human genes and genetic disorders
  - ➔ 11,000+ genes (known sequences) and 6,000+ phenotypes
  - ➔ 500,000+ phenotype-GO associations, including 33,000 genes from 10 species
- **Genecards** (www.genecards.org)
  - ➔ a compendium of genes, protein and diseases
  - ➔ tools to integrate 70+ sources (also OMIM) to a location for info of 24,000+ genes with relationships to diseases
- **Swissprot** (www.ebi.ac.uk/swissprot)
  - ➔ A database of protein sequences with disease annotations for 2600 of its 270,000 entries (16,600 for human proteins)

Databases with disease annotation

- **PhenomicDB** (www.phenomicDB.de)
  - ➔ phenotype-genotype database integrating data from multiple organisms (human and others)
- **Gene2Disease** (www.ogic.ca/projects/g2d_2)
  - ➔ assigns properties to genes related to diseases
  - ➔ provides list of candidates by PubMed MeSH terms and GO
- **GAO** (Genetic Association Database: http://geneticassociationdb.nih.gov)
  - ➔ identify medically relevant polymorphism from the large volume of polymorphism and mutational data
- **Kegg disease** (www.genome.jp/kegg/disease)
  - ➔ genetic & genomic information resource for human diseases

Outline

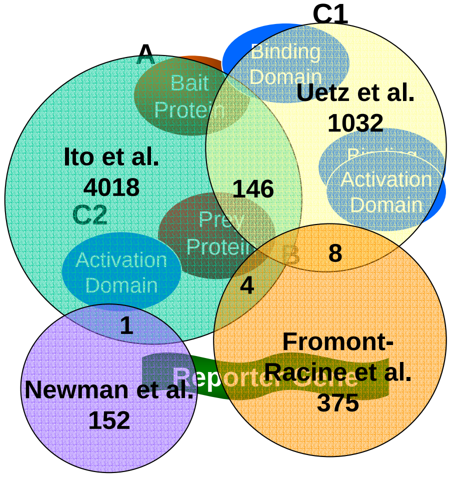
Protein networks in disease

Prediction of protein-protein interactions

PPI-based prediction of disease genes

Experimental approach to PPI

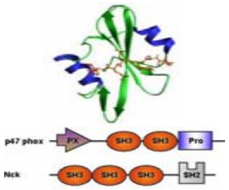
- Many experimental methods for detecting PPIs
  - ➔ Yeast two-hybrid [Ito 01], phage display [Smith 85], mass spectrometry [Bauer 03], etc.
- Limitations of experimental methods
  - ➔ Tedious, labor-intensive
  - ➔ High false positive, high false negative rates
  - ➔ Low consensus among PPI databases



Computational approaches to PPI

Finding rules to say if two given proteins interact?

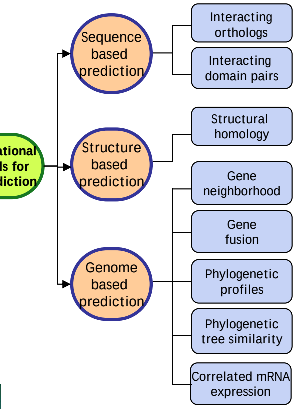
(P_A, P_B) → P_A interacts with P_B?



Domain-based approach

Our work: domain-based approach + multiple database approach

Single database approach


Computational methods for PPI prediction
Sequence based prediction
Interacting orthologs
Interacting domain pairs
Structure based prediction
Structural homology
Genome based prediction
Gene neighborhood
Gene fusion
Phylogenetic profiles
Phylogenetic tree similarity
Correlated mRNA expression

the Gene Ontology
InterDom
Database of Interacting Domains
Pfam
prosite
InterPro
UniProt
the universal protein resource
mips
munich information center for protein sequences
PRINTS
Protein Fingerprint Database

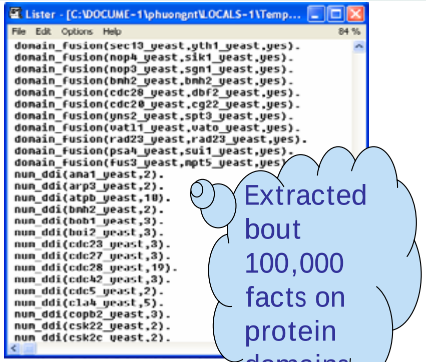
Multiple databases approach

Method: Inductive logic programming

- Input:
  - ➔ Positive examples of interaction pairs:
    - ⇨ Ex: (YAL025C, YJR044C), (YAL026C, YPL146C), …
  - ➔ Negative examples of non-interaction pairs
    - ⇨ Ex: (YAL032C, YLR345W), (YAL032C, YLR 432C), …
  - ➔ Data about protein properties in form of predicates
    - ⇨ Ex: Subcell_cat(YAL025C, cytoplasm), …
- Output: Rules to predict PPI

Extracting protein's domain data

- Extracting domain fusion data from domain fusion database and domain-domain interaction data from iPFAM database.
- Two principle domain features:
  - ➔ Domain fusion
  - ➔ Domain-domain interaction



```
domain_fusion(sec13_yeast,yth1_yeast,yes).
domain_fusion(nop4_yeast,sik1_yeast,yes).
domain_fusion(nop3_yeast,sgn1_yeast,yes).
domain_fusion(bmh2_yeast,bmh2_yeast,yes).
domain_fusion(cdc28_yeast,dbf2_yeast,yes).
domain_fusion(cdc20_yeast,cg22_yeast,yes).
domain_fusion(yns2_yeast,spt3_yeast,yes).
domain_fusion(vat11_yeast,vato_yeast,yes).
domain_fusion(rad23_yeast,rad23_yeast,yes).
domain_fusion(psa4_yeast,sui1_yeast,yes).
domain_fusion(fus3_yeast,mpt5_yeast,yes).
num_ddi(ama1_yeast,2).
num_ddi(arp3_yeast,2).
num_ddi(atpb_yeast,10).
num_ddi(bmh2_yeast,2).
num_ddi(bob1_yeast,3).
num_ddi(boi2_yeast,3).
num_ddi(cdc23_yeast,3).
num_ddi(cdc27_yeast,3).
num_ddi(cdc28_yeast,19).
num_ddi(cdc42_yeast,3).
num_ddi(cdc5_yeast,2).
num_ddi(cla4_yeast,5).
num_ddi(copb2_yeast,3).
num_ddi(csk22_yeast,2).
num_ddi(csk2c_yeast,2).
```

Extracted bout 100,000 facts on protein domains.

Domain fusion

domain fusion(+protein, +protein, #FUSION): A protein pairs has a domain fusion

Domain-domain interaction

hasddi(+protein, +protein, #DDI): A protein pairs has a domain-domain interaction

num ddi(+protein, #NUM DDI): A protein has the number of domain-domain interaction

Extracting genomic/proteomic data from multi databases

Exploiting genomic/proteomic ground facts about proteins and protein interactions from multiple databases.

the Gene Ontology
InterDom
Database of Interacting Domains
prosite
InterPro
Pfam
UniProt
the universal protein resource
PRINTS
Protein Fingerprint Database
mips
munich information center for protein sequences

Extracting genomic/proteomic data from multi databases

For each p ∈ P

SWISS-PROT (protein annotation information)

haskw(+Protein, #Keyword): A protein contains a keyword

hasft(+Protein, #Feature): A protein contains a feature

ec(+Protein, #EC): An enzyme code for a protein

pfam(+Protein, -PFAM_Domain): A protein contains a Pfam

interpro(+Protein, -InterPro_Domain): A protein contains a domain

pir(+Protein, -PIR_Domain): A protein contains a Pir doma

prosite(+Protein, -PROSITE_Domain): A protein contains a domain

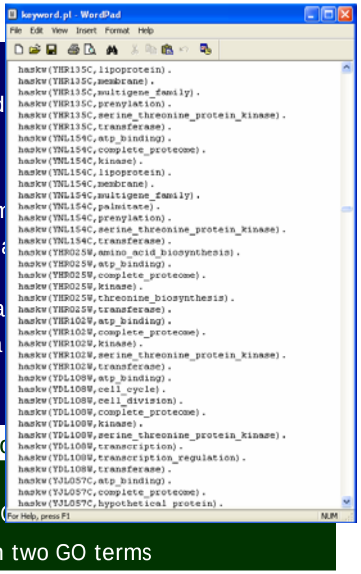
go(+Protein, -GO_Term): A protein contains a GO term

For each (p, q) ∈ G_P x G_P, G_P is set of GO terms asso

GO ("is_a" and "is_part" relations)

is_a(+GO_Term, -GO_Term): is_a relation between two G

part_of(+GO_Term, -GO_Term): part_of relation between two GO terms



```
haskw(YHR135C,lipoprotein).
haskw(YHR135C,membrane).
haskw(YHR135C,multigene_family).
haskw(YHR135C,prenylation).
haskw(YHR135C,serine_threonine_protein_kinase).
haskw(YHR135C,transferase).
haskw(YNL154C,atp_binding).
haskw(YNL154C,complete_proteome).
haskw(YNL154C,kinase).
haskw(YNL154C,lipoprotein).
haskw(YNL154C,membrane).
haskw(YNL154C,multigene_family).
haskw(YNL154C,palmitate).
haskw(YNL154C,prenylation).
haskw(YNL154C,serine_threonine_protein_kinase).
haskw(YNL154C,transferase).
haskw(YHR025W,amino_acid_biosynthesis).
haskw(YHR025W,atp_binding).
haskw(YHR025W,complete_proteome).
haskw(YHR025W,kinase).
haskw(YHR025W,threonine_biosynthesis).
haskw(YHR025W,transferase).
haskw(YHR102W,atp_binding).
haskw(YHR102W,complete_proteome).
haskw(YHR102W,kinase).
haskw(YHR102W,serine_threonine_protein_kinase).
haskw(YHR102W,transferase).
haskw(YDL108W,atp_binding).
haskw(YDL108W,cell_cycle).
haskw(YDL108W,cell_division).
haskw(YDL108W,complete_proteome).
haskw(YDL108W,kinase).
haskw(YDL108W,serine_threonine_protein_kinase).
haskw(YDL108W,transcription).
haskw(YDL108W,transcription_regulation).
haskw(YDL108W,transferase).
haskw(YJL057C,atp_binding).
haskw(YJL057C,complete_proteome).
haskw(YJL057C,hypothetical protein).
```

Extracting genomic/proteomic data from multi databases

For each (p, q) ∈ P x P

Gene Expression (protein expression correlation coefficients)

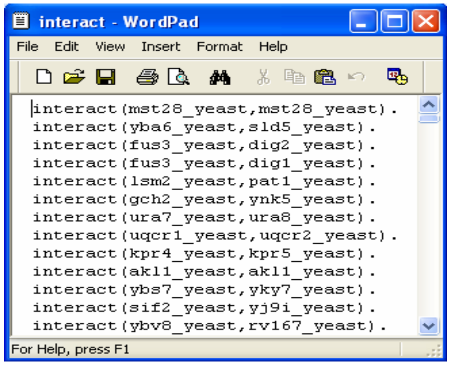
correlation(+Protein, +Protein, -Expression): Expression correlation coefficient between two proteins

For each (p, q) ∈ P x P

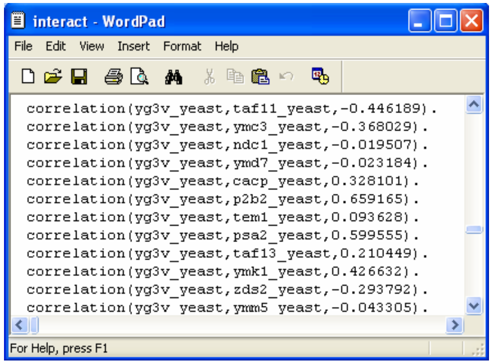
InterPro (protein expression correlation coefficients)

interpro2go(+InterPro_Domain, -GO_Term): Mapping of InterPro entries to GO

Extracted more 200,000 genomic and proteomic facts



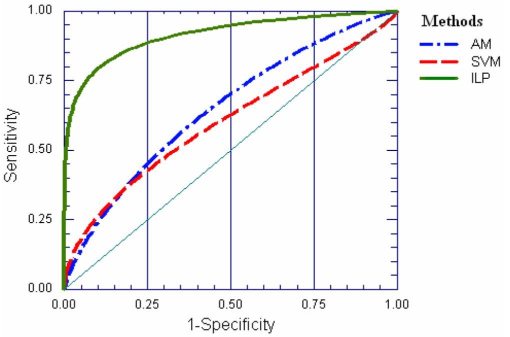
```
interact(mst28_yeast,mst28_yeast).
interact(yba6_yeast,sld5_yeast).
interact(fus3_yeast,dig2_yeast).
interact(fus3_yeast,dig1_yeast).
interact(lsm2_yeast,pat1_yeast).
interact(gch2_yeast,ynk5_yeast).
interact(ura7_yeast,ura8_yeast).
interact(uqcr1_yeast,uqcr2_yeast).
interact(kpr4_yeast,kpr5_yeast).
interact(akl1_yeast,akl1_yeast).
interact(ybs7_yeast,yky7_yeast).
interact(sif2_yeast,yj9i_yeast).
interact(ybv8_yeast,rv167_yeast).
```



```
correlation(yg3v_yeast,taf11_yeast,-0.446189).
correlation(yg3v_yeast,ymc3_yeast,-0.368029).
correlation(yg3v_yeast,ndc1_yeast,-0.019507).
correlation(yg3v_yeast,ymd7_yeast,-0.023184).
correlation(yg3v_yeast,cacp_yeast,0.328101).
correlation(yg3v_yeast,p2b2_yeast,0.659165).
correlation(yg3v_yeast,tem1_yeast,0.093628).
correlation(yg3v_yeast,psa2_yeast,0.599555).
correlation(yg3v_yeast,taf13_yeast,0.210449).
correlation(yg3v_yeast,ymk1_yeast,0.426632).
correlation(yg3v_yeast,zds2_yeast,-0.293792).
correlation(yg3v yeast,ymm5 yeast,-0.043305).
```

Comparison of domain-based methods

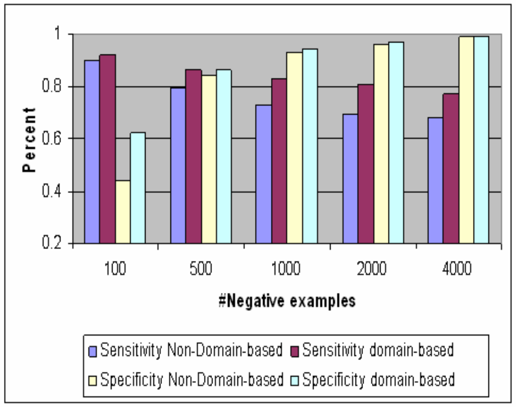
- 5512 positive examples taken from DIP(5963 PPI pairs)
- Negative examples taken in two cases:
  - ➔ With the random negative set: the ROC curve from multiple 10-fold cross validation
  - ➔ With the non co-located negative set: sensitivity and specificity of multiple 10-fold cross validation



	ILP	SVMs	AM
Sensitivity	**84%**	82%	47%
Specificity	**90%**	34%	75%

Comparison of integrative methods

- 10-fold cross validation evaluations for an ILP method with multiple genomic databases, but not using domain features (Tran *et al.*, 2005)
- Our methods performed better with domain features



Some rules obtained

- Rule 1 [Pos cover = 37, Neg cover = 0]

  has_int(A,B) :- subcell cat(B,nucleus),
  subcell cat(A,cytoplasm),
  function_cat(A,transcription).

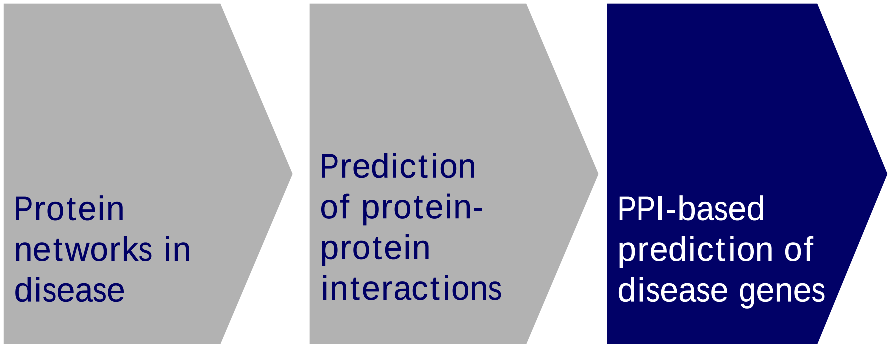
- Rule 2 [Pos cover = 29, Neg cover = 0]

  has_int(A,B) :- ig (A, B, C), C = 1, ddi (A, B, yes),
  function_cat (B, cell rescue defense and virulence).

- Rule 3 [Pos cover = 23, Neg cover = 0]

  interact_domain(A, B) :- go (B, C), is a (C, D),
  hasft (A, chain bud site selection protein bud5).

Outline



Protein networks in disease

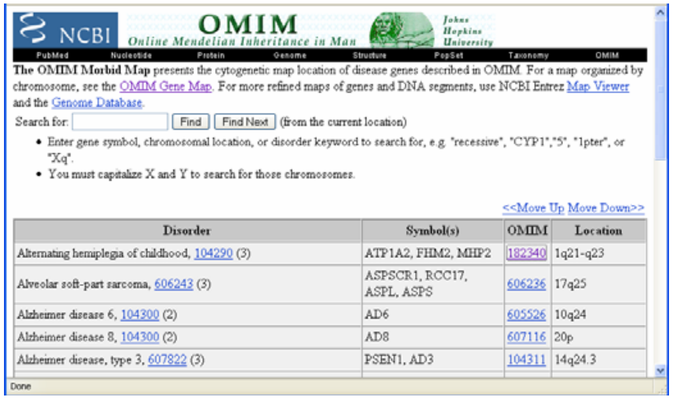
Prediction of protein-protein interactions

PPI-based prediction of disease genes

Disease gene prediction by computation

Problem

- 3,053 already known as disease-causing genes reported in OMIM database (from 25,000-30,000 human genes)
- Predict novel disease-causing genes by computation?



The OMIM Morbid Map presents the cytogenetic map location of disease genes described in OMIM. For a map organized by chromosome, see the OMIM Gene Map. For more refined maps of genes and DNA segments, use NCBI Entrez Map Viewer and the Genome Database.

Search for: [ ] Find Find Next (from the current location)

- Enter gene symbol, chromosomal location, or disorder keyword to search for, e.g. "recessive", "CYP1","5", "1pter", or "Xq".
- You must capitalize X and Y to search for those chromosomes.

<<Move Up Move Down>>

Disorder	Symbol(s)	OMIM	Location
Alternating hemiplegia of childhood, 104290 (3)	ATP1A2, FHM2, MHP2	182340	1q21-q23
Alveolar soft-part sarcoma, 606243 (3)	ASPSCR1, RCC17, ASPL, ASPS	606236	17q25
Alzheimer disease 6, 104300 (2)	AD6	605526	10q24
Alzheimer disease 8, 104300 (2)	AD8	607116	20p
Alzheimer disease, type 3, 607822 (3)	PSEN1, AD3	104311	14q24.3

Done

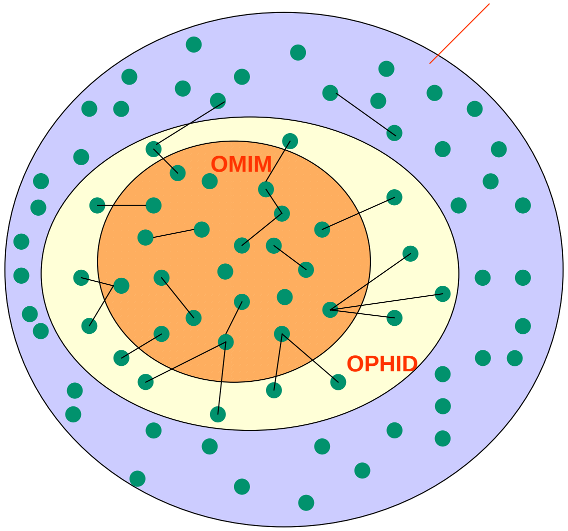
Comp. approaches to disease gene prediction

- Based on annotations (Turner et al., 2003)
- Based on sequences (van Driel et al., 2005)
- Based on protein-protein interactions (PPI)
  - ➔ K-nearest neighbor with PPI data (Xu and Li, 2006)
  - ➔ Heuristic score functions for Alzheimer disease (Chen et al., 2006)
  - ➔ Graph kernels for gene expression and human PPI data (Borgwardt and Kriegel, 2007)
- We developed a new semi-supervised learning (SSL) method based on protein-protein interactions.

Key idea of the method

1. Consider 3590 disease proteins (from OMIM)
2. Consider all interacted proteins from OPHID (51,934 interactions)
3. Consider proteins not belonging to OMIM but interact with OMIM as candidates (5775 cand.)
4. Evaluate the candidates by their score to predict putative disease proteins (found 50 from 5775)

Proteomic/Genomic features from PFAM, GO, UNIPROT, gene expression, Reactome, Interdom

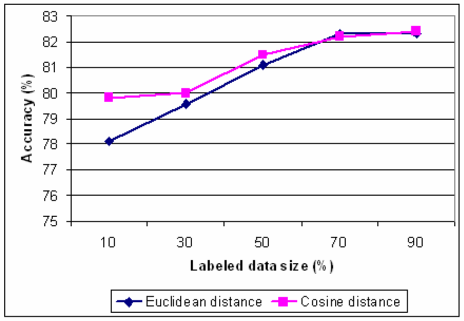


Experiments: data

- Disease proteins: OMIM database (3,053 disease genes) corresponding 3,590 disease proteins.
- Non-disease proteins: Not belong to neither list of ubiquitously expressed human genes (UEHG) nor disease protein data set.
- Candidate disease proteins: 5,775 proteins
- Human PPI data: OPHID database (51,934 interactions)
- Proteomic/Genomic features: Pfam, Uniprot, and GO, Gene Expression, Pathways (Reactome DB), Domain-domain interaction (InterDom DB).

Experiments: results

- We performed 20 trials:
  - ➔ Randomly selected *l* data points as labeled data, and the rest *(n-l)* as unlabeled data.
  - ➔ Estimated accuracy by comparing the predicted labels and true labels.



- Accuracy of our method is from 78% to 82%.
- The recent work of Xu and Li (Bioinformatics 2006) reaches accuracy from 74% to 76%.

Experiments: results

- Implement the k-NN method (using Weka software) on the same data sets
- With various *k* values
- With various scale of training dataset *l*

→ Our method outperformed k-NN method (Xu and Li, 2006)

l % scale of training data set

Method

	10%	30%	50%	70%	90%
K=1	76%	77%	77%	78%	78%
K=3	75%	76%	76%	77%	77%
K=5	74%	75%	75%	76%	76%
K=7	74%	74%	74%	75%	75%
K=9	73%	73%	74%	74%	74%
SSL1	78%	79%	81%	82%	82%
SSL2	80%	80%	81%	82%	82%

* SSL1: SSL method with Euclidean distance

SSL2: SSL method with Cosine distance

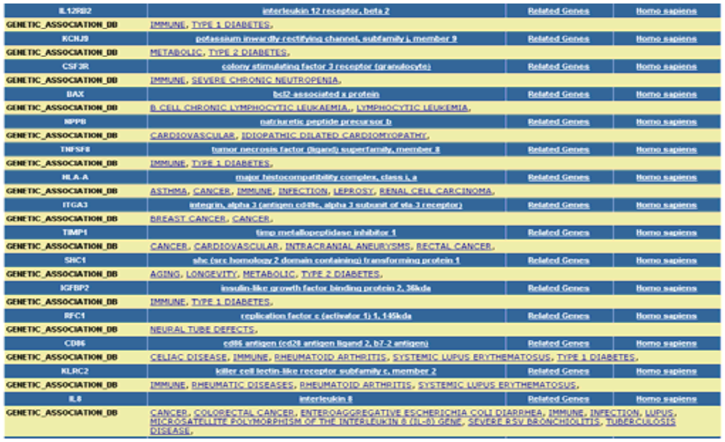
Initial results and interpretation

- We test with all proteins in the human PPI network and newly-predicted 572 disease proteins
- Evaluated indirectly from scientific literatures
  - ➔ Via the function of genes (from databases such as Uniprot, Interpro, GO, etc.) and Medline
  - ➔ Compare with well-known disease gene databases
  - ➔ Via the biological processes such as signal transduction pathways
  - ➔ Via the gene expression

Initial results and interpretation

Among 572 putative proteins (568 GeneID), 29 genes in 67 records found in GAO, e.g.:

- IFNAR1 (interferon alpha, beta and omega receptor 1);
- IGFBP2 (insulin-like growth factor binding protein 2, 36kda);
- TNFSF8 (tumor necrosis factor (ligand) superfamily, member 8).

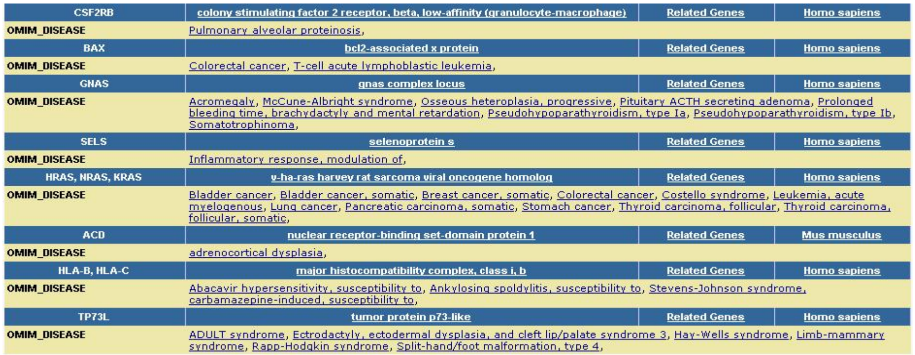


IL12RB2	interleukin 12 receptor, beta 2	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, TYPE 1 DIABETES,		
KCNJ9	potassium inwardly-rectifying channel, subfamily j, member 9	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	METABOLIC, TYPE 2 DIABETES,		
CSF3R	colony stimulating factor 3 receptor (granulocyte)	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, SEVERE CHRONIC NEUTROPENIA,		
BAX	bcl2-associated x protein	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	B CELL CHRONIC LYMPHOCYTIC LEUKAEMIA., LYMPHOCYTIC LEUKEMIA,		
NPPB	natriuretic peptide precursor b	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	CARDIOVASCULAR, IDIOPATHIC DILATED CARDIOMYOPATHY,		
TNFSF8	tumor necrosis factor (ligand) superfamily, member 8	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, TYPE 1 DIABETES,		
HLA-A	major histocompatibility complex, class i, a	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	ASTHMA, CANCER, IMMUNE, INFECTION, LEPROSY, RENAL CELL CARCINOMA,		
ITGA3	integrin, alpha 3 (antigen cd49c, alpha 3 subunit of vla-3 receptor)	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	BREAST CANCER, CANCER,		
TIMP1	timp metallopeptidase inhibitor 1	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	CANCER, CARDIOVASCULAR, INTRACRANIAL ANEURYSMS, RECTAL CANCER,		
SHC1	shc (src homology 2 domain containing) transforming protein 1	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	AGING, LONGEVITY, METABOLIC, TYPE 2 DIABETES,		
IGFBP2	insulin-like growth factor binding protein 2, 36kda	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, TYPE 1 DIABETES,		
RFC1	replication factor c (activator 1) 1, 145kda	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	NEURAL TUBE DEFECTS,		
CD86	cd86 antigen (cd28 antigen ligand 2, b7-2 antigen)	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	CELIAC DISEASE, IMMUNE, RHEUMATOID ARTHRITIS, SYSTEMIC LUPUS ERYTHEMATOSUS, TYPE 1 DIABETES,		
KLRC2	killer cell lectin-like receptor subfamily c, member 2	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, RHEUMATIC DISEASES, RHEUMATOID ARTHRITIS, SYSTEMIC LUPUS ERYTHEMATOSUS,		
IL8	interleukin 8	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	CANCER, COLORECTAL CANCER, ENTEROAGGREGATIVE ESCHERICHIA COLI DIARRHEA, IMMUNE, INFECTION, LUPUS, MICROSATELLITE POLYMORPHISM OF THE INTERLEUKIN 8 (IL-8) GENE, SEVERE RSV BRONCHIOLITIS, TUBERCULOSIS DISEASE,		

http://david.abcc.ncifcrf.gov/

Initial results and interpretation

Among 572 putative proteins (568 GeneID), 2 genes related to 8 records found in OMIM with terms “Colorectal cancer”: BAX (bcl2-associated x protein) and HRAS, NRAS, KRAS (v-ha-ras harvey rat sarcoma viral oncogene homolog)



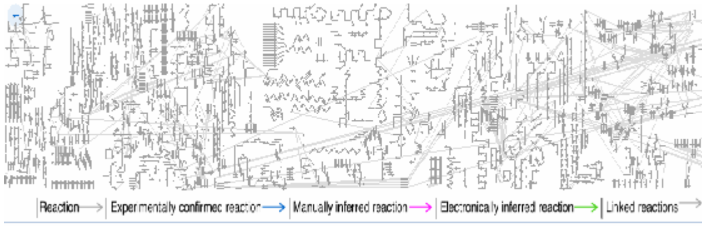
CSF2RB	colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)	Related Genes	Homo sapiens
OMIM_DISEASE	Pulmonary alveolar proteinosis,		
BAX	bcl2-associated x protein	Related Genes	Homo sapiens
OMIM_DISEASE	Colorectal cancer, T-cell acute lymphoblastic leukemia,		
GNAS	gnas complex locus	Related Genes	Homo sapiens
OMIM_DISEASE	Acromegaly, McCune-Albright syndrome, Osseous heteroplasia, progressive, Pituitary ACTH secreting adenoma, Prolonged bleeding time, brachydactyly and mental retardation, Pseudohypoparathyroidism, type Ia, Pseudohypoparathyroidism, type Ib, Somatotrophinoma,		
SELS	selenoprotein s	Related Genes	Homo sapiens
OMIM_DISEASE	Inflammatory response, modulation of,		
HRAS, NRAS, KRAS	v-ha-ras harvey rat sarcoma viral oncogene homolog	Related Genes	Homo sapiens
OMIM_DISEASE	Bladder cancer, Bladder cancer, somatic, Breast cancer, somatic, Colorectal cancer, Costello syndrome, Leukemia, acute myelogenous, Lung cancer, Pancreatic carcinoma, somatic, Stomach cancer, Thyroid carcinoma, follicular, Thyroid carcinoma, follicular, somatic,		
ACD	nuclear receptor-binding set-domain protein 1	Related Genes	Mus musculus
OMIM_DISEASE	adrenocortical dysplasia,		
HLA-B, HLA-C	major histocompatibility complex, class i, b	Related Genes	Homo sapiens
OMIM_DISEASE	Abacavir hypersensitivity, susceptibility to, Ankylosing spoldylitis, susceptibility to, Stevens-Johnson syndrome, carbamazepine-induced, susceptibility to,		
TP73L	tumor protein p73-like	Related Genes	Homo sapiens
OMIM_DISEASE	ADULT syndrome, Ectrodactyly, ectodermal dysplasia, and cleft lip/palate syndrome 3, Hay-Wells syndrome, Limb-mammary syndrome, Rapp-Hodgkin syndrome, Split-hand/foot malformation, type 4,		

http://david.abcc.ncifcrf.gov/

Initial results and interpretation

572 putative proteins sharing

- 47 Reactome pathways with known disease proteins:
  - ➔ Signaling in Immune system (29 putative proteins/74 known DP/103 proteins), e.g. O00459, P01112, P04439
  - ➔ Hemostasis (25 putative proteins), e.g. O00459, P01112, P04085
  - ➔ Gene Expression pathways (21 putative proteins), e.g. O60563
- 270 common UNIPROT keywords with known disease proteins (alternative_splicing (212 proteins), polymorphism (195 proteins), glycoprotein (187 proteins)).

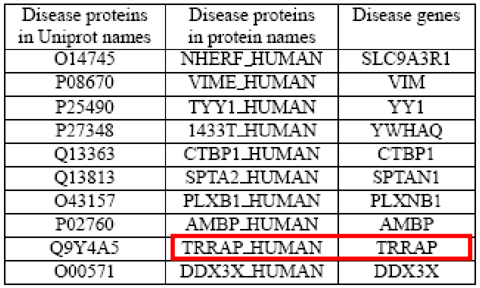

Reaction
Experimentally confirmed reaction
Manually inferred reaction
Electronically inferred reaction
Linked reactions

Initial results and interpretation

Expression of transdominant mutants of the protein trrap human or antisense RNA blocks c-Myc and E1A-mediated oncogenic transformation.

→ **TRRAP** was suggested as an essential cofactor for both the c-Myc and E1A/E2F oncogenic transcription factor pathways.

Table 2. List of some potential disease proteins and corresponding disease genes.

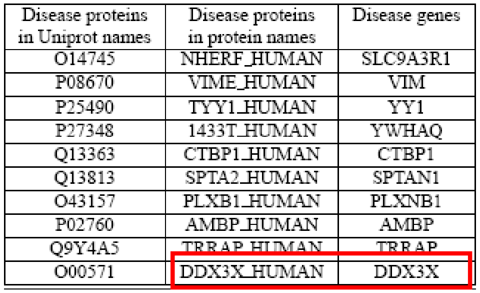


Disease proteins in Uniprot names	Disease proteins in protein names	Disease genes
O14745	NHERF_HUMAN	SLC9A3R1
P08670	VIME_HUMAN	VIM
P25490	TYY1_HUMAN	YY1
P27348	1433T_HUMAN	YWHAQ
Q13363	CTBP1_HUMAN	CTBP1
Q13813	SPTA2_HUMAN	SPTAN1
O43157	PLXB1_HUMAN	PLXNB1
P02760	AMBP_HUMAN	AMBP
Q9Y4A5	TRRAP_HUMAN	TRRAP
O00571	DDX3X_HUMAN	DDX3X

Initial results and interpretation

- **DDX3X** human*:*
  - ➔ Acts as a cofactor for XPO1-mediated nuclear export of incompletely spliced HIV-1 Rev RNAs
  - ➔ Is involved in HIV-1 replication.
- Protein HIV-1 interacts specifically with hepatitis C virus core protein (Owsianka, 1999).

Table 2. List of some potential disease proteins and corresponding disease genes.

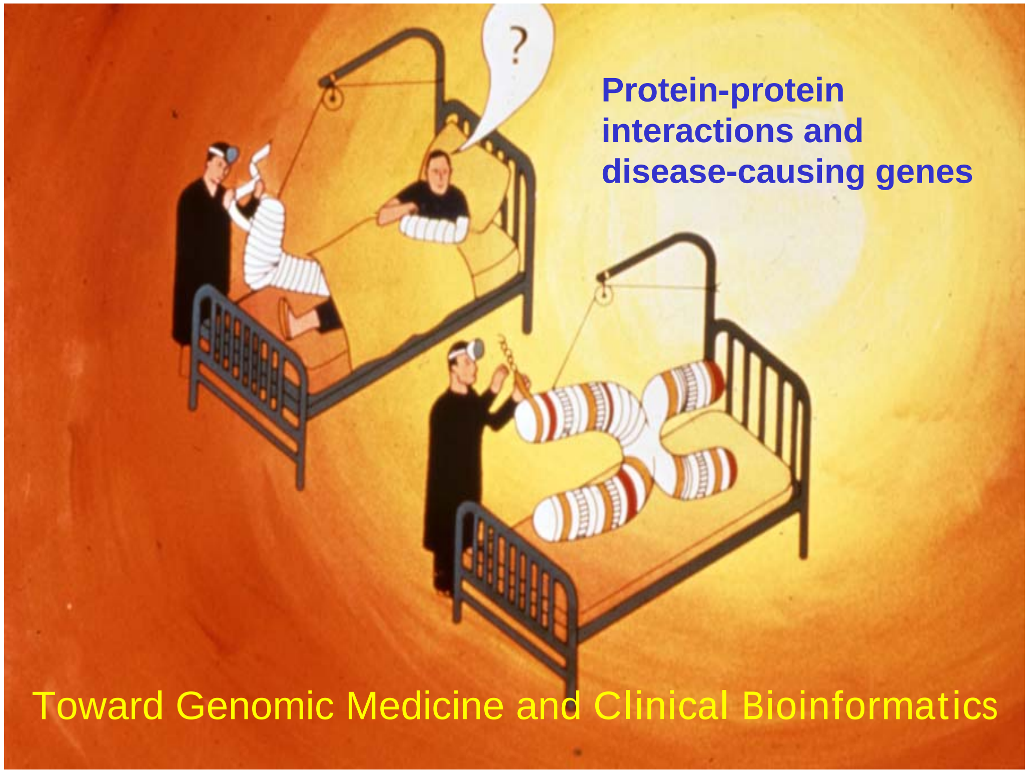


Disease proteins in Uniprot names	Disease proteins in protein names	Disease genes
O14745	NHERF_HUMAN	SLC9A3R1
P08670	VIME_HUMAN	VIM
P25490	TYY1_HUMAN	YY1
P27348	1433T_HUMAN	YWHAQ
Q13363	CTBP1_HUMAN	CTBP1
Q13813	SPTA2_HUMAN	SPTAN1
O43157	PLXB1_HUMAN	PLXNB1
P02760	AMBP_HUMAN	AMBP
Q9Y4A5	TRRAP_HUMAN	TRRAP
O00571	DDX3X_HUMAN	DDX3X

→ DDX3X should be a candidate of hepatitis C disease genes.

Conclusion

- Large protein network databases are now available and have an increasing importance in disease study.
- Computational methods allow us to exploit them.
- Our preliminary work in prediction of protein-protein interactions and disease-causing genes
- Look towards a joint research: similarity measure evaluation of putative genes, potential features, clinical data for disease gene prediction, etc.



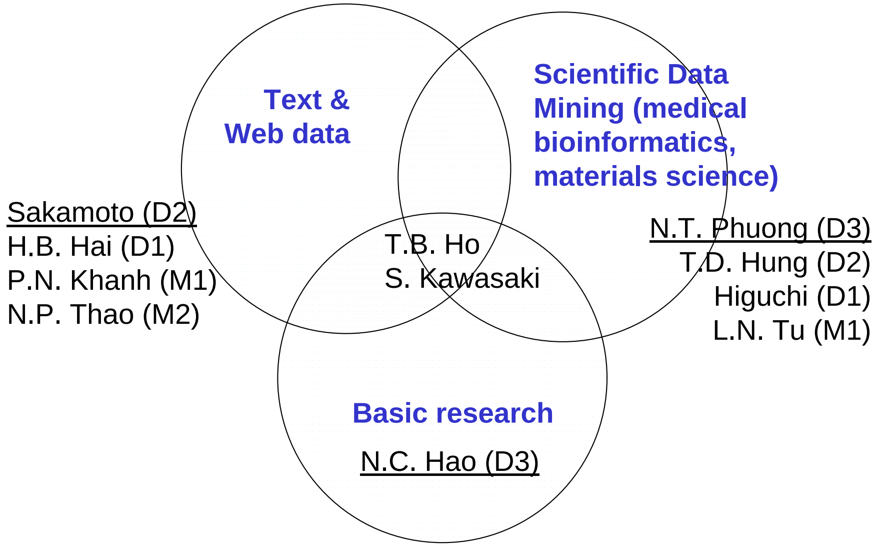
Protein-protein interactions and disease-causing genes

Toward Genomic Medicine and Clinical Bioinformatics

Key references

- Ideker,T., Sharan, R., Protein networks in disease, *Genome Research*, April 2008, 1-9.
- Kann, M.G. 2007. Protein interactions and disease: Computational approaches to uncover the etiology of diseases. *Brief. Bioinform.* **8:** 333–346.
- Nguyen, T.P., Ho, T.B.: Combining Domain Fusions and Domain-Domain Interactions to Predict Protein-Protein Interactions, *Journal of Bioinformatics and Computational Biology* (accepted).
- Nguyen, T.P., Ho, T.B.: A Semi-Supervised Learning Approach to Disease Gene Prediction, *IEEE International Conference on BioInformation and BioMedicine (BIBM'07)*, Silicon Valley, 423-428, November 2-4, 2007.

Knowledge creating methodology lab



Text & Web data

Scientific Data Mining (medical bioinformatics, materials science)

Basic research

Sakamoto (D2)
H.B. Hai (D1)
P.N. Khanh (M1)
N.P. Thao (M2)

T.B. Ho
S. Kawasaki

N.T. Phuong (D3)
T.D. Hung (D2)
Higuchi (D1)
L.N. Tu (M1)

N.C. Hao (D3)

Experiment design

- Comparative experiments to validate:
  - ➔ the advantages of the integration of multiple proteomic and genomic features.
  - ➔ the advantages of domain-based approach.
- Experiments
  - ➔ 10 times of 10-fold cross validation to compare with domain-based methods, i.e., AM (Sprinzak et al. 2001]) and SVM (SVMlight)
  - ➔ 10 times of 10-fold cross validation to compare with integrative methods, i.e., ILP (Tran *et al.*, 2005)

Experiments: design

- Evaluate the computational performance of the proposed semi-supervised learning method
  - ➔ Multiple tests with different parameters to calculate accuracy of the proposed method
  - ➔ Compare with a supervised learning method, k-nearest neighbor (Xu and Li, *Bioinformatics* 2006)
- Verify new putative disease genes
  - ➔ Investigate scientific literature to look for evidences