Development and Application of Protein-Protein interaction Prediction System, PreDIN (Prediction-oriented Database of Interaction Network)

서 정근 박사 (LGCI Life Science R&D 책임연구원)
(042-866-2297, 019-9154-1266, suhjung@lgci.co.kr)

Abstract:

Motivation: Protein-protein interaction plays a critical role in the biological processes. The identification of interacting proteins by bioinformatical methods can provide new lead in the functional studies of uncharacterized proteins without performing extensive experiments.

Results: Protein-protein interactions are predicted by a computational algorithm based on the weighted scoring system for domain interactions between interacting protein pairs. Here we propose potential interaction domain (PID) pairs can be extracted from a data set of experimentally identified interacting protein pairs, where one protein contains a domain and its interacting protein contains the other. Every combinations of PID are summarized in a matrix table termed the PID matrix, and this matrix has proposed to be used for prediction of interactions. The database of interacting proteins (DIP) has used as a source of interacting protein pairs and InterPro, an integrated database of protein families, domains and functional sites, has used for defining domains in interacting pairs. A statistical scoring system, named "PID matrix score" has designed and applied as a measure of interaction probability between domains. Cross-validation has been performed with subsets of DIP data to evaluate the prediction accuracy of PID matrix. The prediction system gives about 50% of sensitivity and 98% of specificity. Based on the PID matrix, we develop a system providing several interaction information-finding services in the Internet. The system, named PreDIN (Prediction-oriented Database of Interaction Network) provides interacting domain finding services and interacting protein finding services. It is demonstrated that mapping of the genome-wide interaction network can be achieved by using the PreDIN system. This system can be also used as a new tool for functional prediction of unknown proteins.

약력

LGCI Life Science R&D 책임연구원

2002년도 제1차 한국생물정보학회 워크샵

Computational Aspects of Bioinformatics

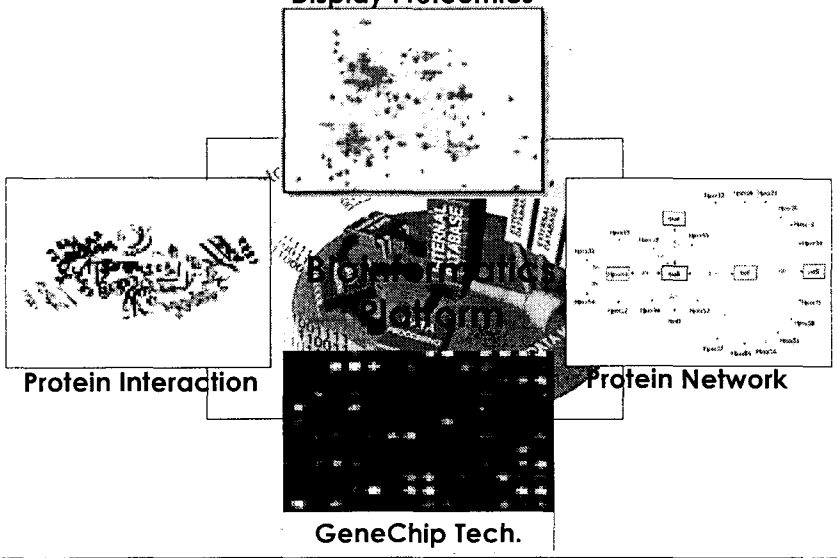
Development and Application of Protein-Protein interaction Prediction System, PreDIN (Prediction-oriented Database of Interaction Network)

LGCI 생명과학기술연구원
바이오텍연구소 **Proteomics Program:** 서 정근

2002. 2. 8.

LG Proteomics Program Overview

LGCI / 생명과학연구원

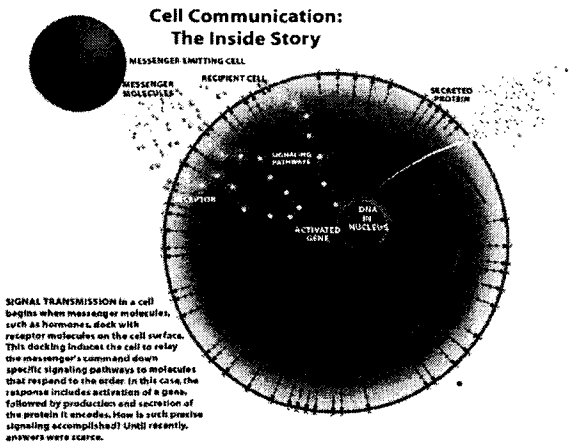


Physiome: Cell Communication

LGCI / 생명과학연구원

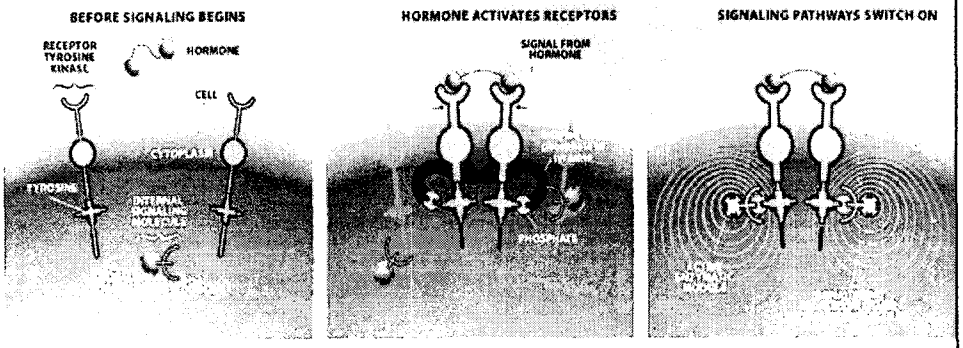
Physiome: 생체 분자 사이의 물리적 상호작용의 총체

- 단백질-단백질
- 단백질-리간드
- 리셉터-리간드
- 항 원-항 체
- 효 소-기 질
- 단백질-DNA
- 단백질-Lipid
- 단백질-Sugar
-
-
-



단백질 상호작용의 중요성

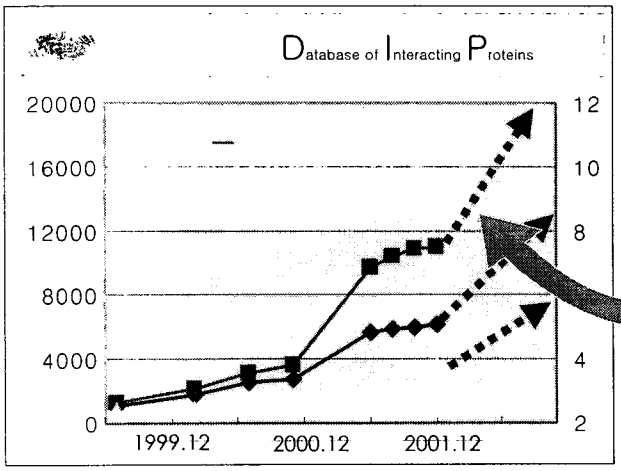
LGCI / 생명과학연구원



Scott and Pawson. Sci. Am. 282,72-79

단백질 상호작용 데이터의 축적

LGCI / 생명과학연구원



On-Going Y2H

Human

Arabidopsis

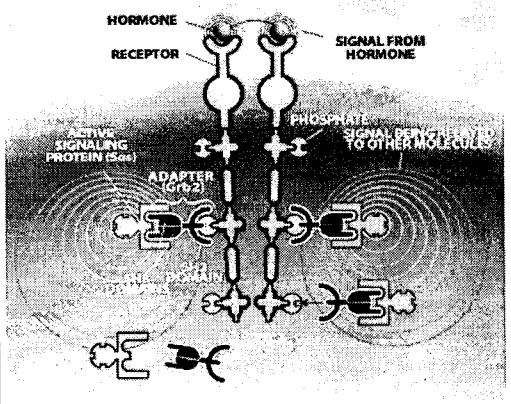
Max-Planck Institut

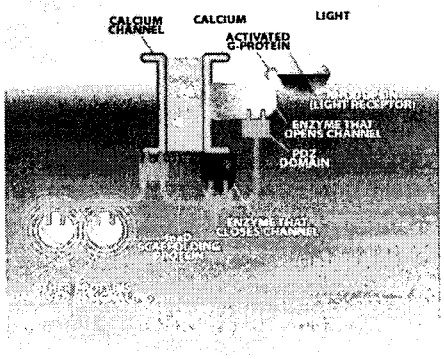
Pathogens

UNIVERSITY OF WASHINGTON

도메인의 중요성

LGCI / 생명과학연구원

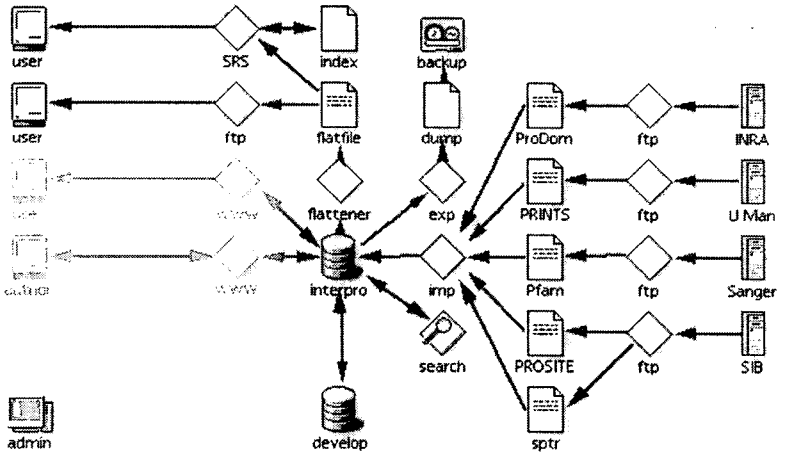




Scott and Pawson. Sci. Am. 282,72-79

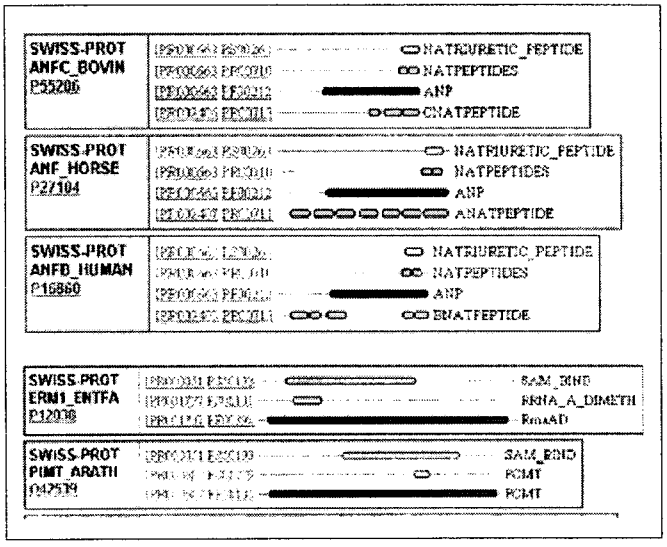
Domain Analysis: InterPro

LGCI / 생명과학연구원



InterPro: Results

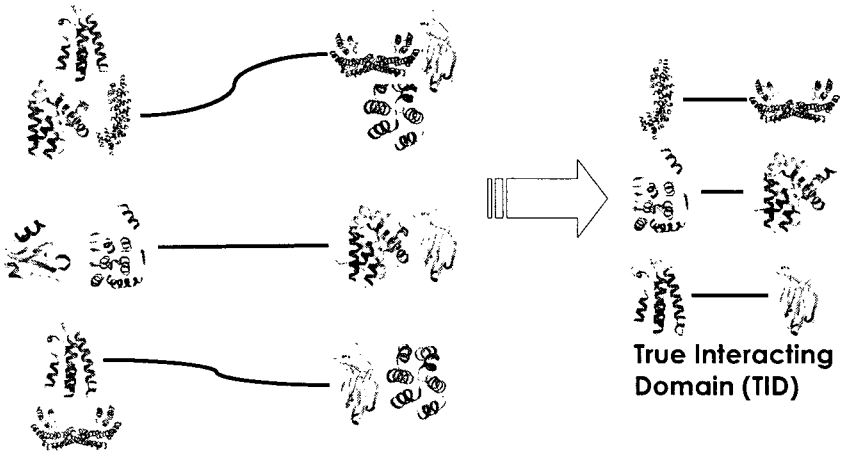
LGCI / 생명과학연구원



도메인 정보의 한계

LGCI / 생명과학연구원

Protein A ---Interaction---Protein B



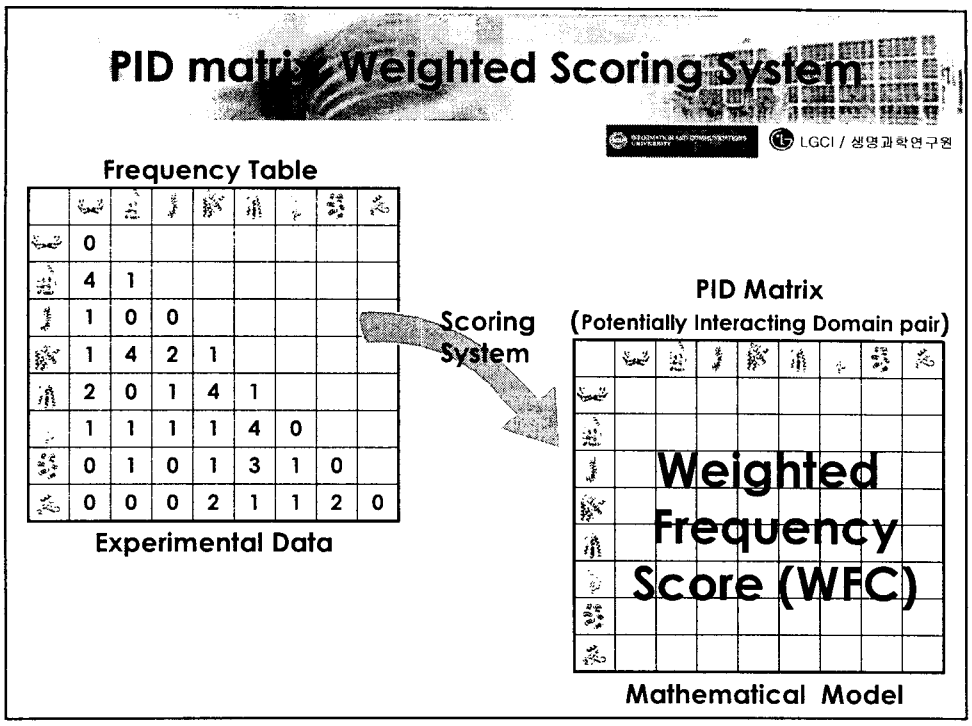
Extracting Domain-Interaction Pattern

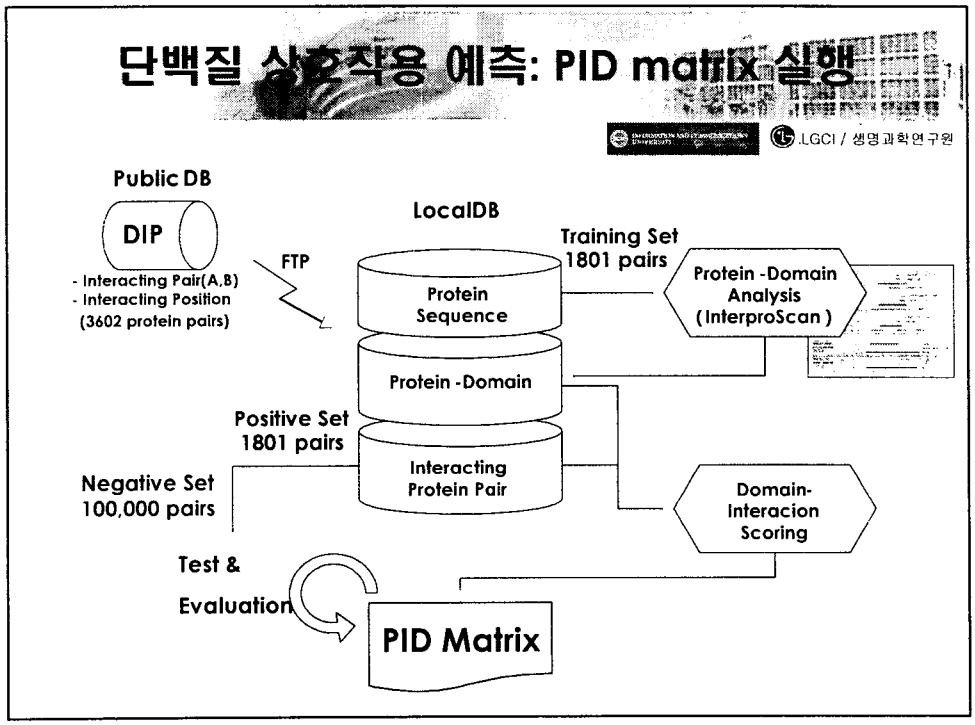
LGCI / 생명과학연구원

	0							
	4	1						
	1	0	0					
	1	4	2	1				
	2	0	1	4	1			
	1	1	1	1	4	0		
	0	1	0	4	3	1	0	
	0	0	0	2	1	1	2	0

Data Set

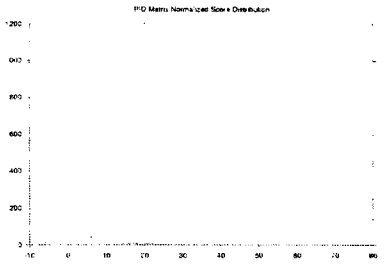
Interaction Pattern (Finding TID)


PID matrix Weighted Scoring System
LGCI / 생명과학연구원
Frequency Table
Experimental Data
Scoring System
PID Matrix
(Potentially Interacting Domain pair)
Weighted Frequency Score (WFC)
Mathematical Model

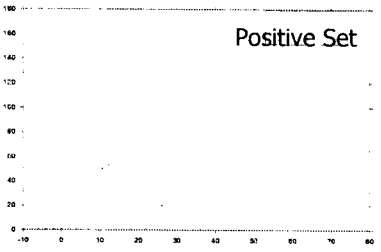

단백질 상호작용 예측: PID matrix 실행
LGCI / 생명과학연구원
Public DB
DIP
- Interacting Pair(A,B)
- Interacting Position
(3602 protein pairs)
FTP
LocalDB
Protein Sequence
Protein -Domain
Interacting Protein Pair
Training Set 1801 pairs
Protein -Domain Analysis (InterproScan)
Positive Set 1801 pairs
Negative Set 100,000 pairs
Domain-Interacion Scoring
Test & Evaluation
PID Matrix

단백질 상호작용 예측 시스템 검증

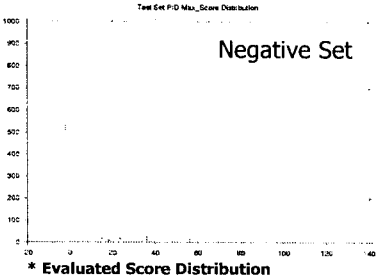
LGCI / 생명과학연구원



PID matrix



Positive Set



Negative Set

* Evaluated Score Distribution

단백질 상호작용 예측 시스템 검증 결과

LGCI / 생명과학연구원

	Positive Set (DIP database)	Negative Set (Embl/SwissProt)
전체 단백질 쌍의 수 (a)	1801	100000
도메인을 포함하는 단백질 쌍의 수 (b)	1304	100000
상호 작용할 것으로 예측된 단백질 쌍의 수 (c)	832	712
도메인을 포함하는 단백질 쌍의 비율 (%) (b/a × 100 %)	72.9	100.0
도메인을 포함하는 단백질 쌍 중에서 상호 작용할 것으로 예측된 단백질 쌍의 비율 (%) (c/b × 100 %)	63.4 (민감도 1)	0.7
전체 단백질 쌍 중에서 상호 작용할 것으로 예측된 단백질 쌍의 비율 (%) (c/a × 100 %)	46.2 (민감도 2)	0.7

Comparisons

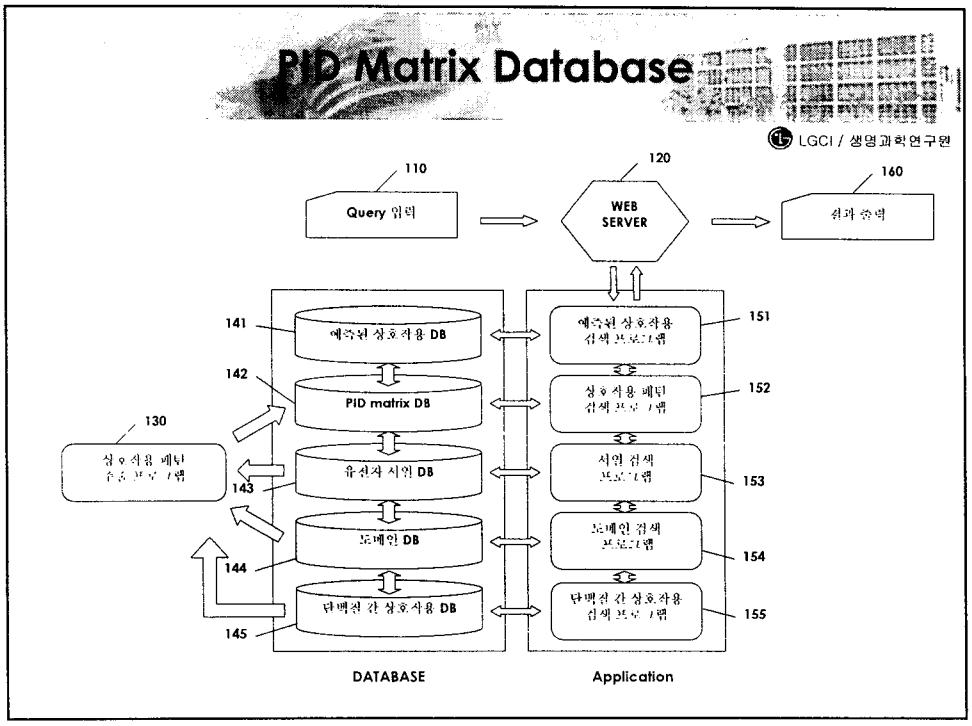
LGCI / 생명과학연구원

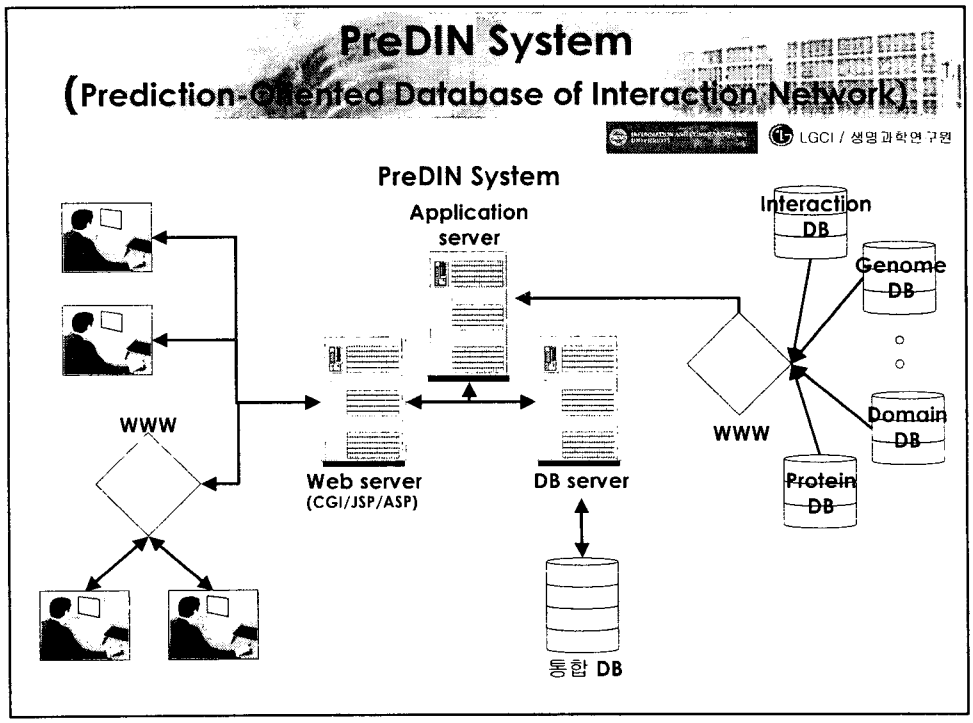
	PID Matrix	SVM	Interacting Domain Profile	Combined Algorithm
Training Set	1/2 of DIP (3602)	1/2 of DIP(2664)	1524 (PIM)	Experimental : 500 (DIP,MIPS) Phylogenic Profile : 20749 mRNA expression : 26013 Rosetta Stone : 45502 Metabolic Function :2391
Positive Test Set	1/2 of DIP (3602)	1/2 of DIP(2664)	E.coli proteome	Swiss-Prot (keywords)
Random Test Set	Random Pair Embl/SwissProt	k-let Shuffled Sequence		
Sensitivity (TP/(TP+FP))	46.2%	80 % (accuracy)	-	55.6%
Specificity (TF/(TF+FF))	> 99%		-	> 83.9%
Comment	Domain 포함해야 적용 (약 60~70%)	Generally Applicable	Not evaluated with +/- dataset	Functional Links Prediction

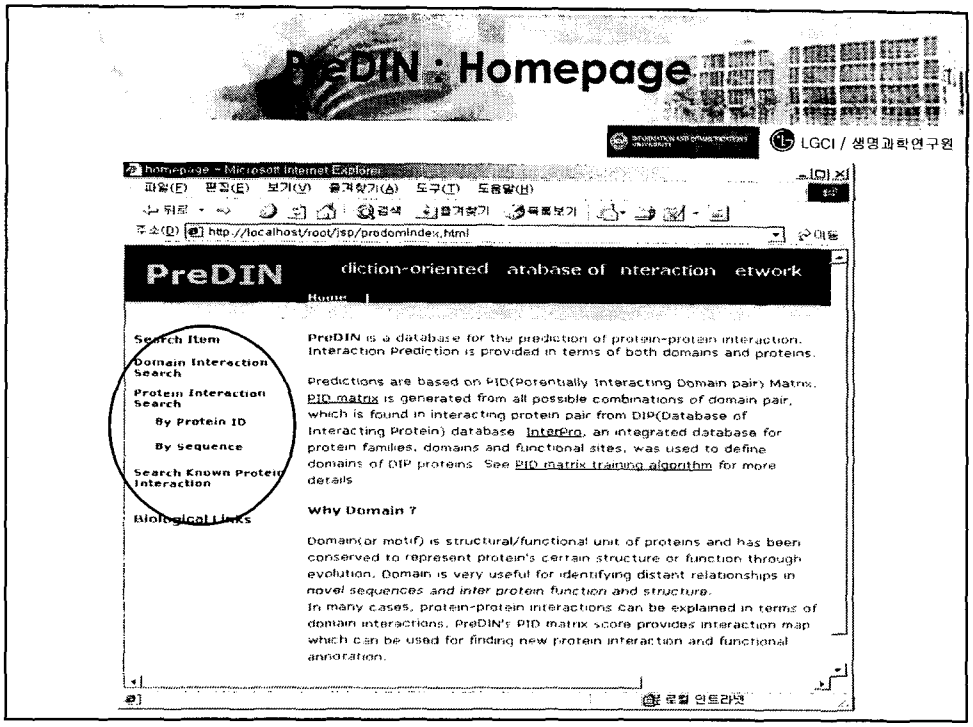
Hot Spot Domains in PID matrix

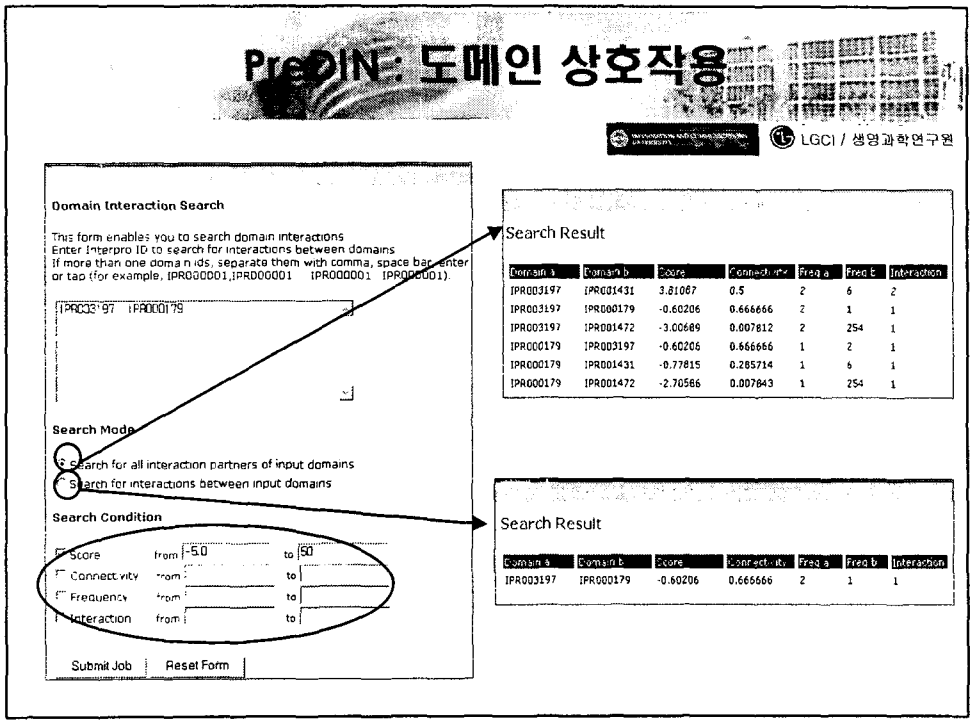
LGCI / 생명과학연구원

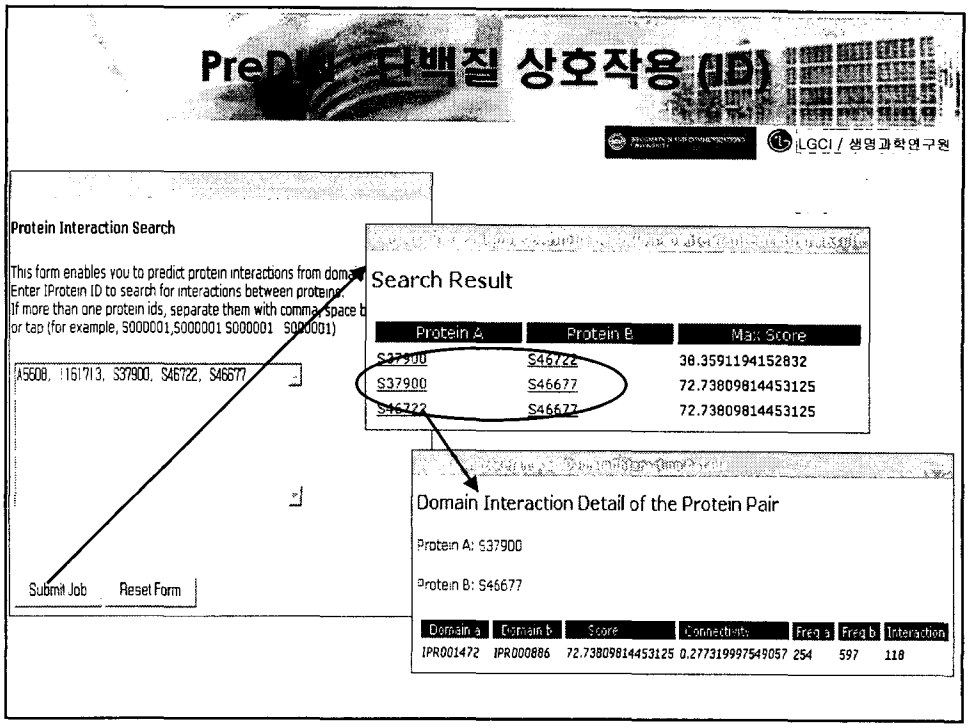
Domain	+ Partner	Doman	Domain	+ Partner	Doman
IPR000095	30	PAK-box /P21-Rho-binding	IPR001650	33	Helicase C-terminal domain
IPR000108	21	Neutrophil cytosol factor 2	IPR001660	25	SAM domain (Sterile alpha motif)
IPR000166	29	Histone-fold/TFIID-TAF/NF-Y domain	IPR001680	65	G-protein beta WD-40 repeats
IPR000194	38	ATP synthase alpha and beta subunit, N-terminal	IPR001683	21	PX (Bem1/NCF1/PI3K) domain
IPR000225	29	Armadillo repeat	IPR001687	209	ATP/GTP-binding site motif A (P-loop)
IPR000345	34	Cytochrome c family heme-binding site	IPR001806	69	Ras GTPase superfamily
IPR000488	23	Death domain	IPR001841	25	RING fin
IPR000504	70	RNA-binding region RNP-1 (RNA recognition motif)	IPR001849	33	Pleckstrin homology (PH) domain
IPR000536	21	Ligand-binding domain of nuclear hormone receptor	IPR001871	38	bZIP (Basic-leucine zipper) transcription factor family
IPR000553	31	Cyclin	IPR002041	21	GTP-binding nuclear protein Ran family
IPR000561	23	EGF-like domain	IPR002048	33	EF-hand family
IPR000694	42	Proline-rich region	IPR002162	23	D-isomer specific 2-hydroxyacid dehydrogenase
IPR000719	122	Eukaryotic protein kinase	IPR002290	121	Serine/Threonine protein kinase family active site
IPR000727	21	Target SNARE coiled-coil domain	IPR002652	28	Importin beta binding domain
IPR000822	27	Zinc finger, C2H2 type	IPR002965	38	Proline rich extensin
IPR000886	426	**Endoplasmic reticulum targeting sequence**	IPR002996	24	Cytokine receptor class 1 family-specific domain B
IPR000934	23	Serine/threonine specific protein phosphatase	IPR003015	30	Myc-type, helix-loop-helix dimerization domain
IPR000980	45	Src homology 2 (SH2) domain	IPR003527	22	MAP kin
IPR001023	22	Heat shock protein hsp70	IPR003577	61	RAS small GTPases, Ras subfamily
IPR001060	23	Cell division control protein 15 (CDC15)	IPR003578	62	RAS small GTPases, Rho subfamily
IPR001092	22	Helix-loop-helix dimerization domain	IPR003579	62	RAS small GTPases, Rab subfamily
IPR001138	25	Fungal transcriptional regulatory protein, N-terminus	IPR003593	52	AAA ATPase superfamily
IPR001163	41	Small nuclear ribonucleoprotein (Sm protein)	IPR003594	22	Histidine kinase-like ATPase
IPR001230	41	Prenyl group binding site (CAAX box)	IPR003961	23	Fibronectin type III domain
IPR001245	117	Tyrosine kinase catalytic domain	IPR004000	21	Actin and actin-like
IPR001313	40	Pumilio-family RNA binding domains (aka PUM-HD, Pumilio homology domain)	PS50311	31	Cysteine-rich region
IPR001404	21	Heat shock hsp90 proteins family	PS50312	24	Asp-rich region
IPR001410	34	DEAD/DEAH box helicase	PS50315	54	Glycine-rich region
IPR001452	81	Src homology 3 (SH3) domain	PS50318	27	Lysine-rich region
IPR001472	198	Bipartite nuclear localization signal	PS50321	64	Asparagine-rich region
IPR001494	21	Importin-beta N-terminal domain	PS50322	83	Glutamine rich region
IPR001628	21	C4-type steroid receptor zinc finger	PS50324	84	Ser rich

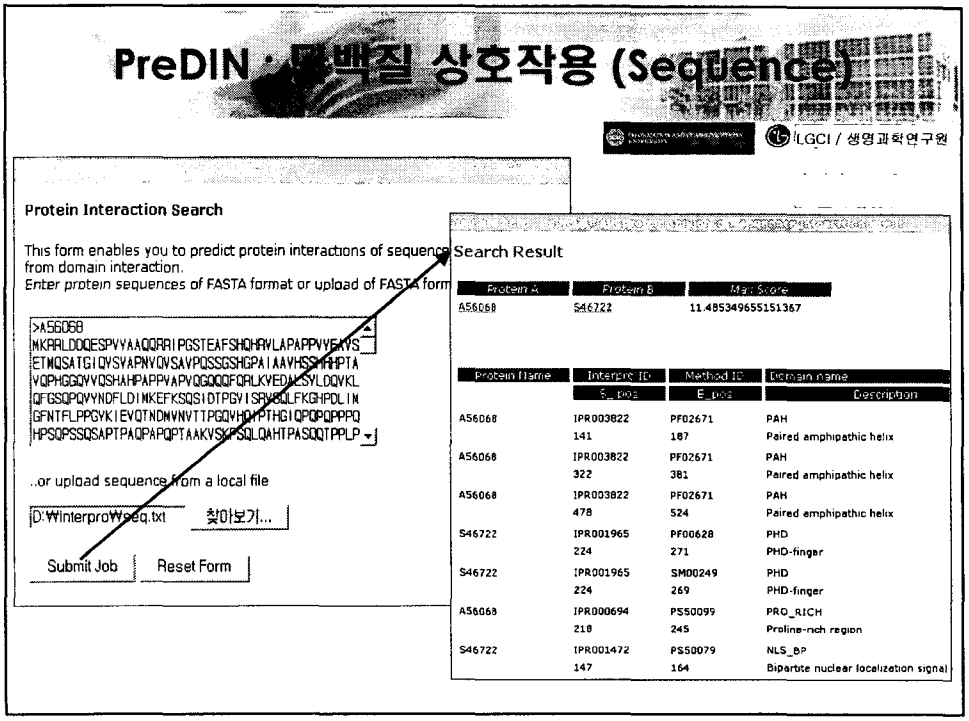

PID Matrix Database
LGCI / 생명과학연구원
110
Query 입력
120
WEB SERVER
160
141
142
130
143
144
145
151
152
153
154
155
PID matrix DB
DATABASE
Application

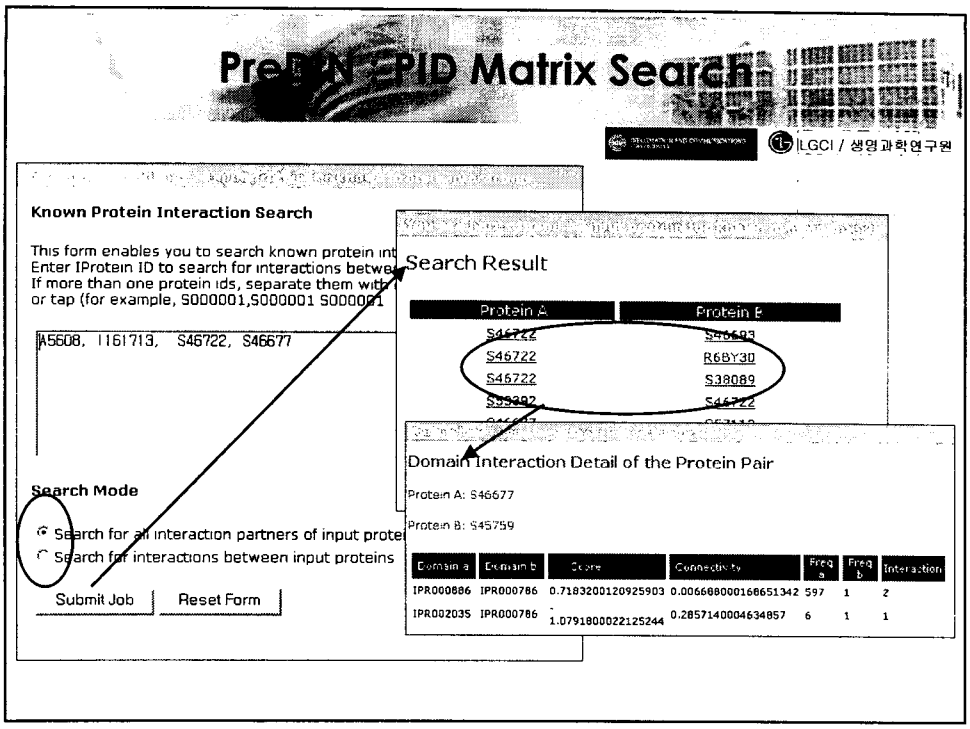

PreDIN System
(Prediction-Oriented Database of Interaction Network)
LGCI / 생명과학연구원
PreDIN System
Application server
Interaction DB
Genome DB
Domain DB
Protein DB
WWW
WWW
Web server
(CGI/JSP/ASP)
DB server
통합 DB

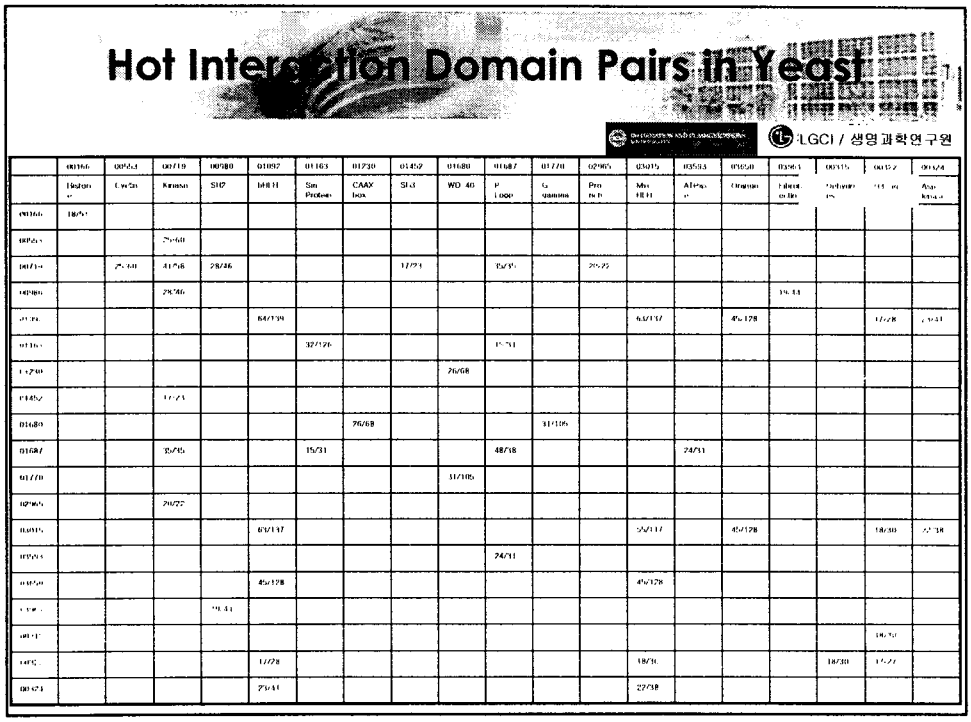

PreDIN : Homepage
LGCI / 생명과학연구원
homepage - Microsoft Internet Explorer
http://localhost/root/jsp/predomIndex.html
PreDIN diction-oriented atabase of nteraction etwork
Home
Search Item
Domain Interaction Search
Protein Interaction Search
By Protein ID
By Sequence
Search Known Protein Interaction
Biological Links
PreDIN is a database for the prediction of protein-protein interaction. Interaction Prediction is provided in terms of both domains and proteins.
Predictions are based on PID(Potentially Interacting Domain pair) Matrix. PID matrix is generated from all possible combinations of domain pair, which is found in interacting protein pair from DIP(Database of Interacting Protein) database. InterPro, an integrated database for protein families, domains and functional sites, was used to define domains of DIP proteins. See PID matrix training algorithm for more details
Why Domain ?
Domain(or motif) is structural/functional unit of proteins and has been conserved to represent protein's certain structure or function through evolution. Domain is very useful for identifying distant relationships in novel sequences and infer protein function and structure.
In many cases, protein-protein interactions can be explained in terms of domain interactions. PreDIN's PID matrix score provides interaction map which can be used for finding new protein interaction and functional annotation.


PreDIN: 도메인 상호작용
LGCI / 생명과학연구원
Domain Interaction Search
This form enables you to search domain interactions
Enter Interpro ID to search for interactions between domains
If more than one domain ids, separate them with comma, space bar, enter or tab (for example, IPR000001,IPR000001 IPR000001 IPR000001).
IPR003197 IPR000179
Search Mode
Search for all interaction partners of input domains
Search for interactions between input domains
Search Condition
Score from -5.0 to 50
Connectivity from to
Frequency from to
Interaction from to
Submit Job
Reset Form
Search Result
Domain a | Domain b | Score | Connectivity | Freq a | Freq b | Interaction
IPR003197 | IPR001431 | 3.81067 | 0.5 | 2 | 6 | 2
IPR003197 | IPR000179 | -0.60206 | 0.666666 | 2 | 1 | 1
IPR003197 | IPR001472 | -3.00689 | 0.007812 | 2 | 254 | 1
IPR000179 | IPR003197 | -0.60206 | 0.666666 | 1 | 2 | 1
IPR000179 | IPR001431 | -0.77815 | 0.285714 | 1 | 6 | 1
IPR000179 | IPR001472 | -2.70586 | 0.007843 | 1 | 254 | 1
Search Result
Domain a | Domain b | Score | Connectivity | Freq a | Freq b | Interaction
IPR003197 | IPR000179 | -0.60206 | 0.666666 | 2 | 1 | 1


LGCI / 생명과학연구원
Protein Interaction Search
Search Result
Protein A
Protein B
Max Score
S37900
S46722
38.3591194152832
S37900
S46677
72.73809814453125
S46722
S46677
72.73809814453125
A5608, 1161713, S37900, S46722, S46677
Submit Job
Reset Form
Domain Interaction Detail of the Protein Pair
Protein A: S37900
Protein B: S46677
Domain a
Domain b
Score
Connectivity
Freq a
Freq b
Interaction
IPR001472
IPR000886
72.73809814453125
0.277319997549057
254
597
118


PreDIN : 단백질 상호작용 (Sequence)
LGCI / 생명과학연구원
Protein Interaction Search
Search Result
Protein A
Protein B
Max Score
A56068
S46722
11.485349655151367
Protein Name
Interpro ID
Method ID
Domain name
S_pos
E_pos
Description
A56068
IPR003822
PF02671
PAH
141
187
Paired amphipathic helix
A56068
IPR003822
PF02671
PAH
322
381
Paired amphipathic helix
A56068
IPR003822
PF02671
PAH
478
524
Paired amphipathic helix
S46722
IPR001965
PF00628
PHD
224
271
PHD-finger
S46722
IPR001965
SM00249
PHD
224
269
PHD-finger
A56068
IPR000694
PS50099
PRO_RICH
218
245
Proline-rich region
S46722
IPR001472
PS50079
NLS_BP
147
164
Bipartite nuclear localization signal
..or upload sequence from a local file
찾아보기...
Submit Job
Reset Form

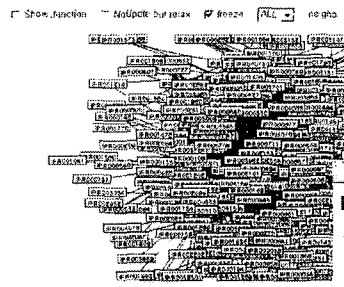

PreDDI - PID Matrix Search
ILGCI / 생명과학연구원
Known Protein Interaction Search
This form enables you to search known protein int
Enter IProtein ID to search for interactions betwe
If more than one protein ids, separate them with
or tap (for example, S000001,S000001 S000001
A5608, I161713, S46722, S46677
Search Mode
Search for all interaction partners of input protei
Search for interactions between input proteins
Submit Job
Reset Form
Search Result
Protein A
Protein B
S46722
R6BY30
S46722
S38089
Domain Interaction Detail of the Protein Pair
Protein A: S46677
Protein B: S45759
Domain a
Domain b
Score
Connectivity
Freq a
Freq b
Interaction
IPR000886
IPR000786
0.7183200120925903
0.006680000168651342
597
1
2
IPR002035
IPR000786
1.0791800022125244
0.2857140004634857
6
1
1


Hot Interaction Domain Pairs in Yeast
LGCI / 생명과학연구원

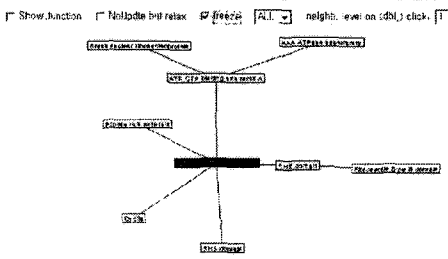
PreDIN Network

LGCI / 생명과학연구원

Yeast Domain



Hot Domain Pairs



Protein Network, Domain Network

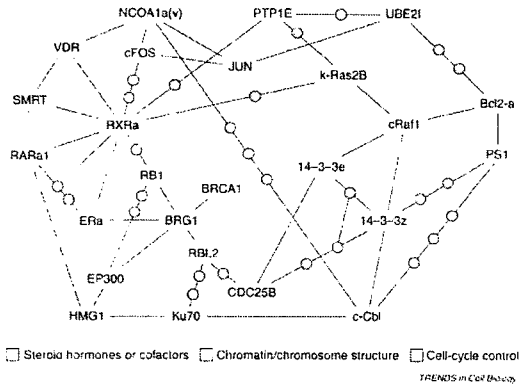
LGCI / 생명과학연구원

Yeast Protein Network

Yeast Domain Network

Yeast Domain Network

LGCI / 생명과학연구원

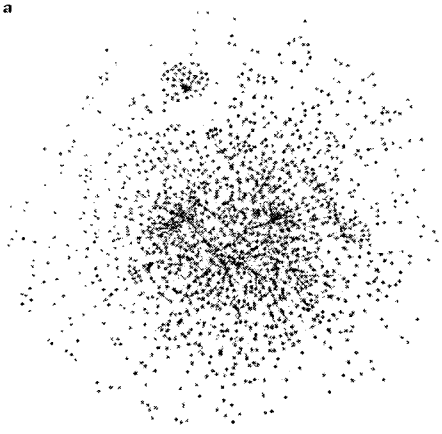


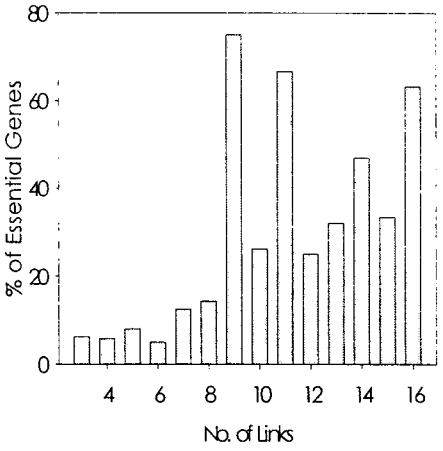
Hot Domain Network

Protein Network: Essential Genes

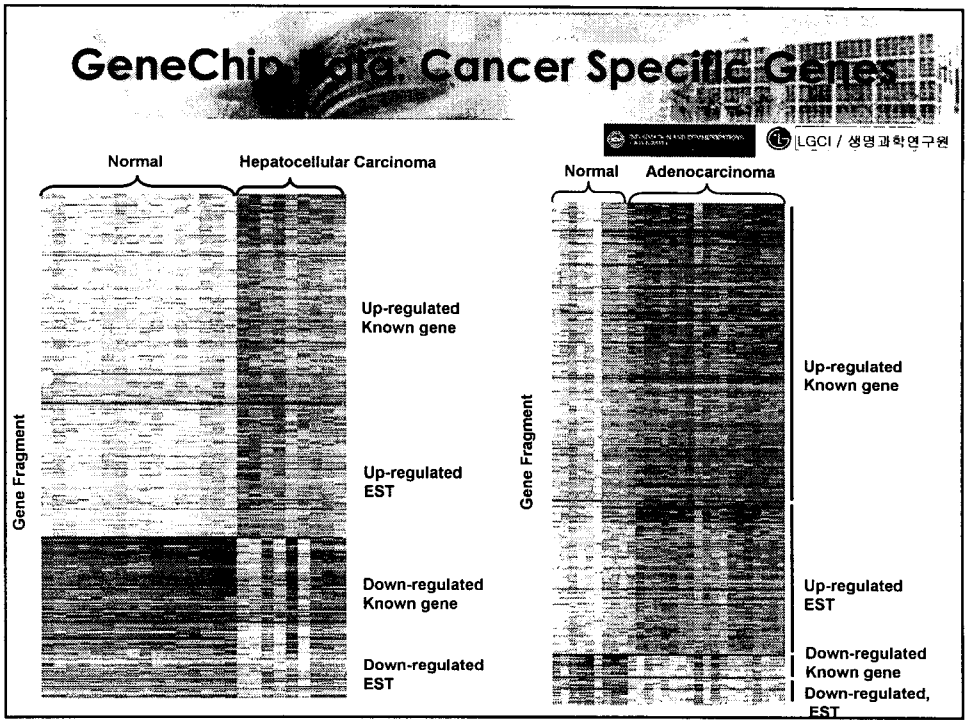
LGCI / 생명과학연구원

a





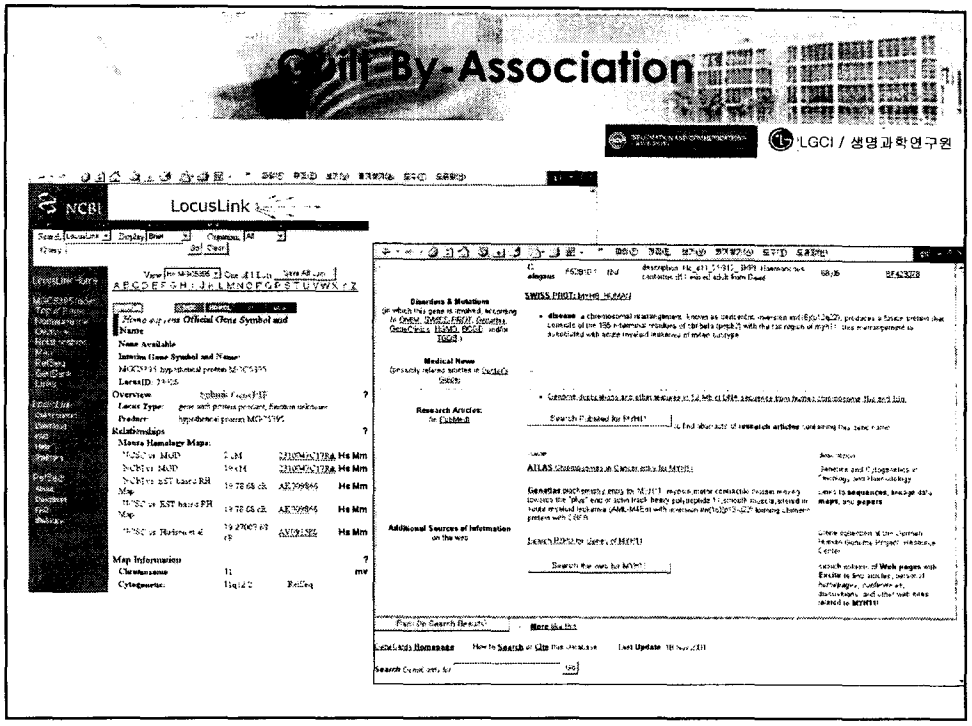
Adapted from Nature 411: 41

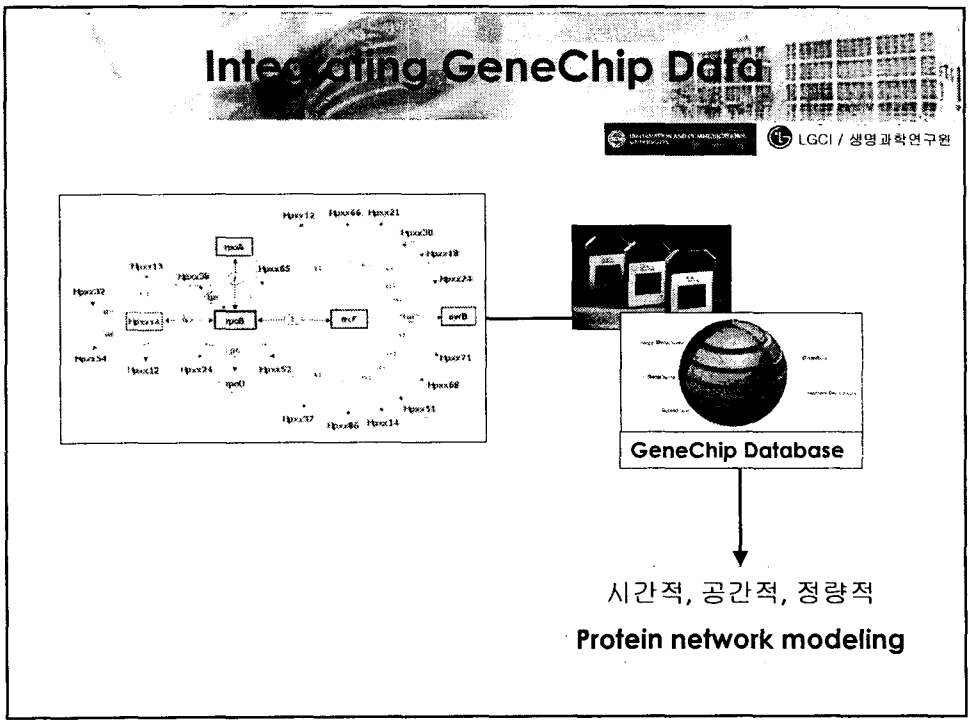


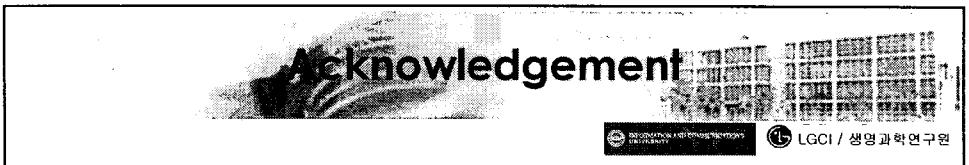
Prioritized Gene List

LGCI / 생명과학연구원

	Known Cancer Association ?		TOTAL GENES	PRIORITIZED GENES
	YES	NO		
Enzyme	108	426	**534**	**180**
Secreted Protein	121	142	**263**	**116**
Transcription Factor	68	99	**167**	**68**
Channel / Transporter	7	119	**126**	**79**
Membrane Protein	79	107	**186**	**81**
Peceptor / G Protein	61	123	**184**	**88**
Translation Factor	0	30	**30**	**0**
Proteosome Component	0	4	**4**	**0**
Chaperone	0	3	**3**	**1**
No Molecular Function Annotated	220	1,529	**1,749**	**0**
TOTAL	**664**	**2,582**	**3,246**	**613**


Guilt-By-Association
LGCI / 생명과학연구원
NCBI
LocusLink


Integrating GeneChip Data
LGCI / 생명과학연구원
GeneChip Database
시간적, 공간적, 정량적
Protein network modeling



LGCI 생명과학 연구원

예측 알고리즘 및 도메인 분석

김 규 완 연구원
김 완 규 연구원
이 은 정 연구원

단백질 분석

지 희 정 박사
천 지 혜 연구원
한 성 연구원

LG-BMI

GeneChip Data 분석

정 현 호 박사
고 상 석 박사
양 두 석 연구원

정보통신대학원 **(ICU)**

DB prototype 개발

한 동 수 박사
김 행 이 연구원