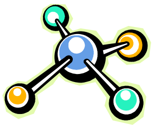
國家衛生研究院 資訊科學研究所

Comparative Genomics and Interactomics



林仲彦[1,2] 熊昭[1]

cylin@nhri.org.tw hsiung@nhri.org.tw

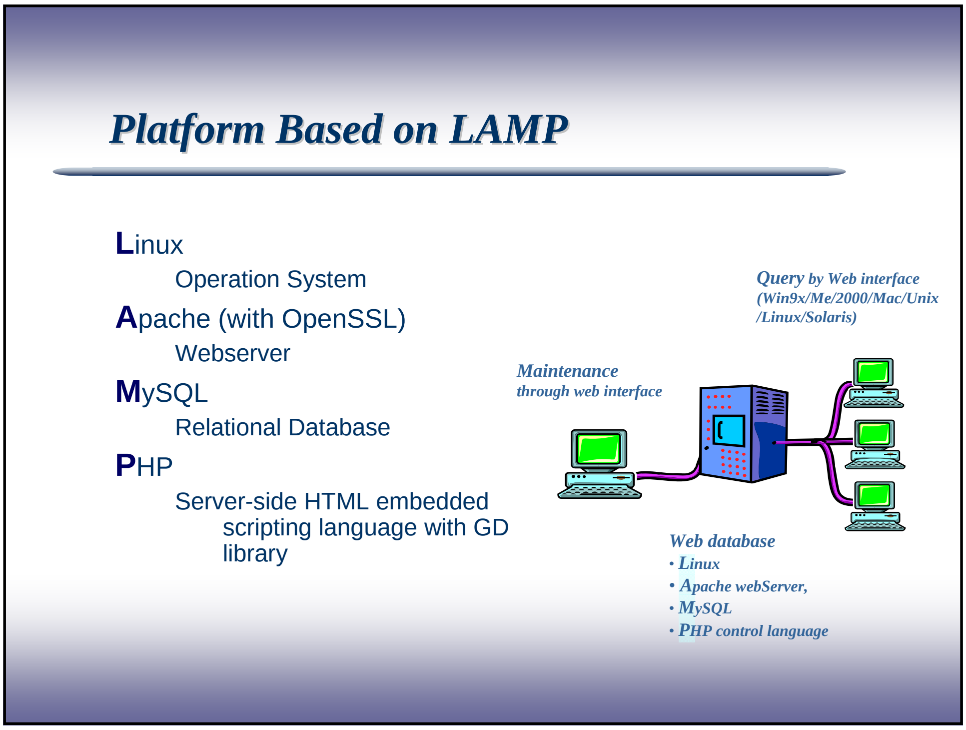
1.Division of Biostatistics and Bioinformatics, NHRI

2.Institute of Information Science, Academia Sinica

July. 2006.

Objectives

- Comparative genomics and Interactomics
- Provide handy bioinformatic tools on-line as listed below:
  - Primer design tool with a web interface.
  - Phylogenetic analysis web tool
  - Protein-protein interaction prediction tool.
- Acquaint users in need with these tools via workshops and assistant services.



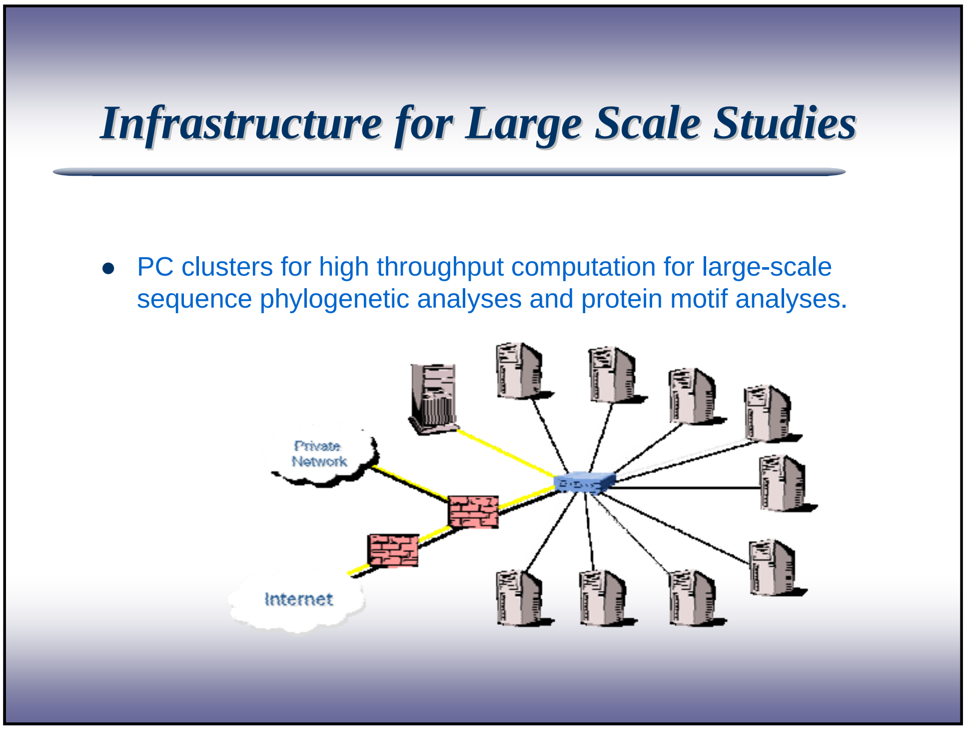
Platform Based on LAMP

Linux
    Operation System
Apache (with OpenSSL)
    Webserver
MySQL
    Relational Database
PHP
    Server-side HTML embedded scripting language with GD library

Maintenance
through web interface

Query by Web interface
(Win9x/Me/2000/Mac/Unix
/Linux/Solaris)

Web database
- *Linux*
- *Apache webServer,*
- *MySQL*
- *PHP control language*



Infrastructure for Large Scale Studies

- PC clusters for high throughput computation for large-scale sequence phylogenetic analyses and protein motif analyses.

Currently available service

- Comparative Genomics
  - Primer Design Assistant (PDA)
    - Assistance of primer design for single or multiple sequences under the same condition
    - Providing a easy way to design the nested primers for specific PCR Reactions

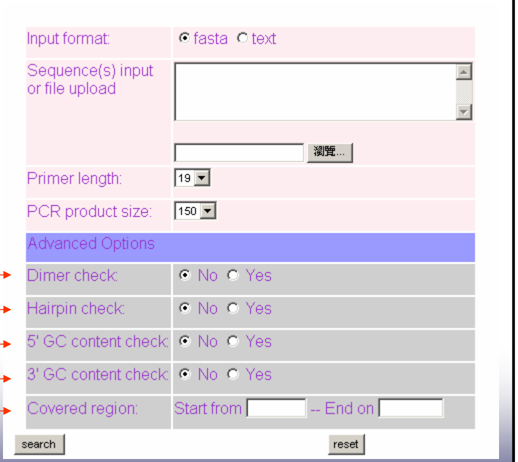
PDA
Primer Design Assistant

http://dbb.nhri.org.tw/primer/

Currently available service (conti)

- Primer Design Assistant (PDA)
  - Customized PCR conditions

Dimer check
Hairpin check
5’GC content check
3’GC content check
Covered region



Input format: ● fasta ○ text

Sequence(s) input or file upload

Primer length: 19

PCR product size: 150

Advanced Options

Dimer check: ● No ○ Yes

Hairpin check: ● No ○ Yes

5' GC content check: ● No ○ Yes

3' GC content check: ● No ○ Yes

Covered region: Start from -- End on

Criterion for PDA Setting

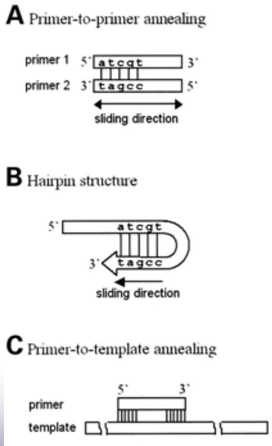
- Default Settings

Repeats	Any four continual nucleotides (AAAA, TTTT, CCCC, or GGGG) will be avoided for both forward and reversed primers. Continuous dinucleotide repeats, such as 'ATATAT', are also avoided.
C/G clamp	G or C on the end of 3' terminal
GC %	25% ~ 75%
Tm	Tm of forward and reversed primers restricted to be higher than 50°C
ΔTm	restricted to be smaller than 5°C

- Advanced options

Dimer check:	This option turns on can avoid primer dimer formation.
Hairpin check:	This option turns on can avoid internal self-complementarity.
5' GC content check:	Check the GC% of 5' to add the ability to recognize the template and enhance the priming specificity.
3' GC content check:	Check the GC% of 3' to avoid mismatch to avoid mismatch.
Covered region:	By entering the start position and stop position, you can get the PCR product containing the segment you need.

Calculation of the Stability of DNA Duplexes


A Primer-to-primer annealing

B Hairpin structure

C Primer-to-template annealing

D Nearest-neighbor parameters for all possible NN dimer duplexes. Modified from SantaLucia 1998 (10).

Sequence	Free Energy Parameter (ΔG°_{50})
A a	-0.73
A t	-0.61
A c	-1.16
A g	-0.92
T a	-0.32
T t	-0.73
T c	-1.03
T g	-1.16
C a	-1.16
C t	-0.92
C c	-1.57
C g	-1.81
G a	-1.03
G t	-1.16
G c	-1.92
G g	-1.57
A	0.98
T	0.98
C	1.00
G	1.00

Ranking Mechanism

The primer pairs passing through the limitations listed above are sorted by ranking score (*R*):

$$R = 100 - \Delta(T_\mathrm{m}) + \Delta G^\circ_\mathrm{forward}(3' - 5') + \Delta G^\circ_\mathrm{reverse}(3' - 5') + \text{hairpin score} + \text{dimer score}$$

Currently available service (conti)

Input format: ● fasta ○ text

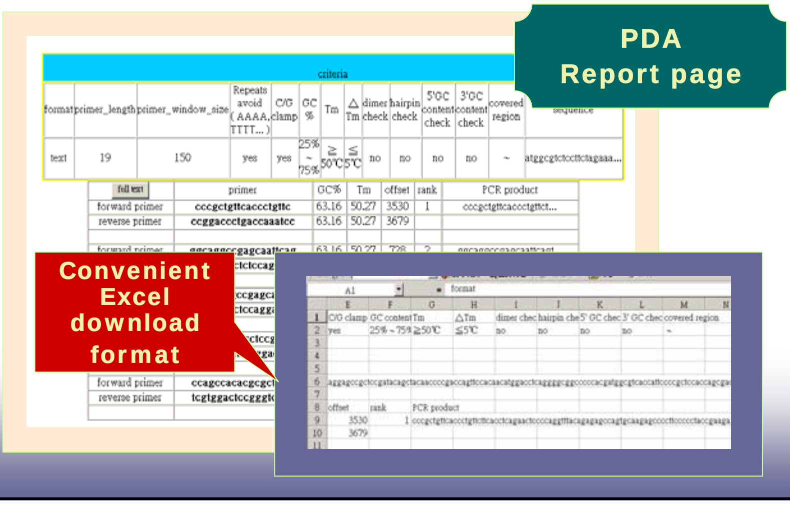
Sequence(s) input or file upload

Primer length: 19

PCR product size: 150

Batch primer design for unified experimental conditions

Currently available service (conti)



PDA Report page

criteria													
format	primer_length	primer_window_size	Repeats avoid (AAAA, TTTT...)	C/G clamp	GC %	Tm	△ Tm	dimer check	hairpin check	5'GC content check	3'GC content check	covered region	sequence
text	19	150	yes	yes	25% ~ 75%	≥ 50℃	≤ 5℃	no	no	no	no	~	atggcgtctccttctagaaa...

full text	primer	GC%	Tm	offset	rank	PCR product
forward primer	**cccgctgttcaccctgttc**	63.16	50.27	3530	1	cccgctgttcaccctgttct...
reverse primer	**ccggaccctgaccaaatcc**	63.16	50.27	3679		

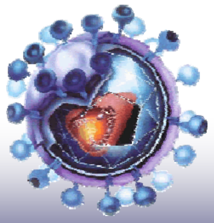
Convenient Excel download format

A1 format

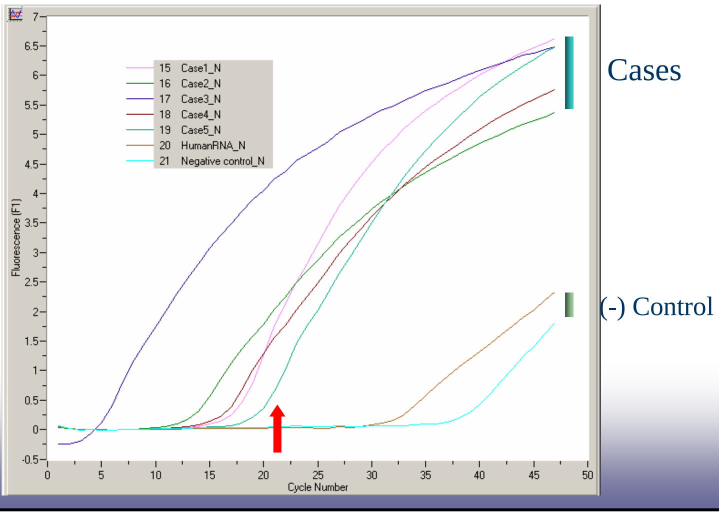
	E	F	G	H	I	J	K	L	M
1	C/G clamp	GC content	Tm	△Tm	dimer chec	hairpin che	5' GC chec	3' GC chec	covered region
2	yes	25% ~ 75%	≥50℃	≤5℃	no	no	no	no	~
6	aggagccgctccgatacagctacaaccccgaccagttccacaacatggacctcagggcggcccccacgatggcgtcaccattcccgctccaccagcga								
8	offset	rank	PCR product						
9	3530	1	cccgctgttcaccctgttcttcacctcagaactcccaggtttacagagagccagtgcaagagcccttcccccctacgaaga						
10	3679								

Use PDA to Develop PCR kits for SARS Detection

- 由於目前通用的SARS-CoV檢驗方法靈敏度有限，使得病毒量較低或是因採樣方法不佳的檢體無法被檢測到，在防疫的前提下，本組便與院內基因醫學組協同疾管局於發展出高靈敏度之檢測方法。
- 檢測方法中所需要的核酸引子都透過PDA來進行設計，避免引子本身dimer 及 hairpins 的形成，加速了檢測方法的建立。
- 此方法為結合1st run RT-PCR + 2nd run Q-PCR，可於1.5小時內檢測出結果，縱使病毒量低於10隻，此一方法也可以檢測出來。

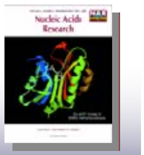


Result of Real-time PCR for SARS-CoV Detection



Primer Design Assistant (PDA)

- Primers designed through PDA has been experimentally proved to reach 97% successful rate
- Free access
- Visits from 2003/07/01~2006/10/16 : ~61,000
- Report primer sequences generated in this period of time: ~360,000



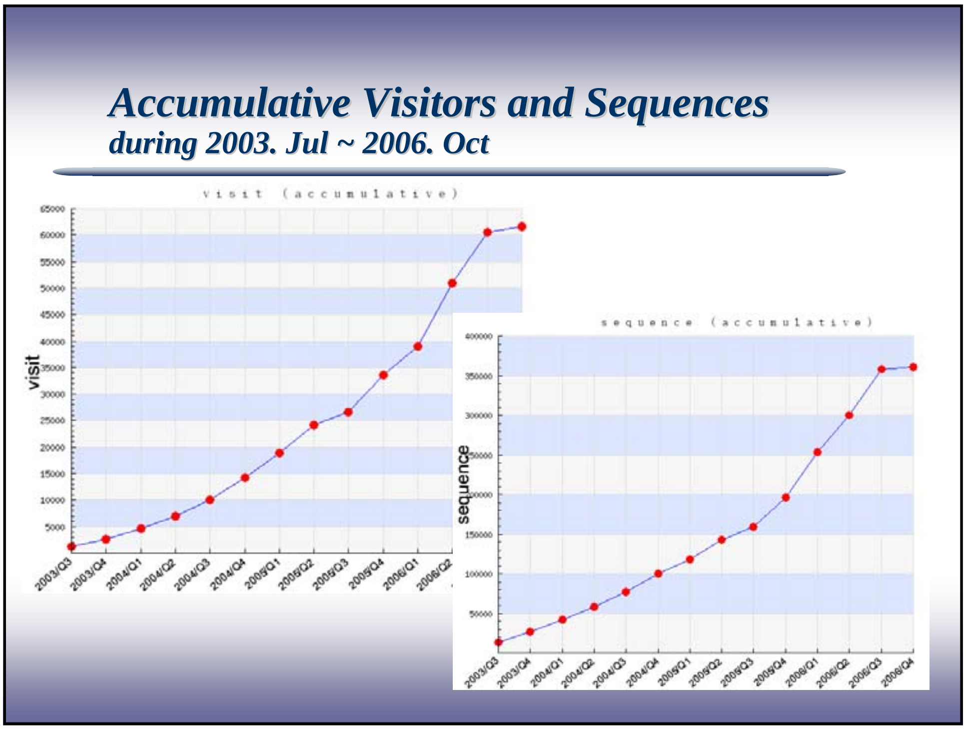
Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3751–3754***
DOI: 10.1093/nar/gkg560

Primer Design Assistant (PDA): a web-based primer design tool

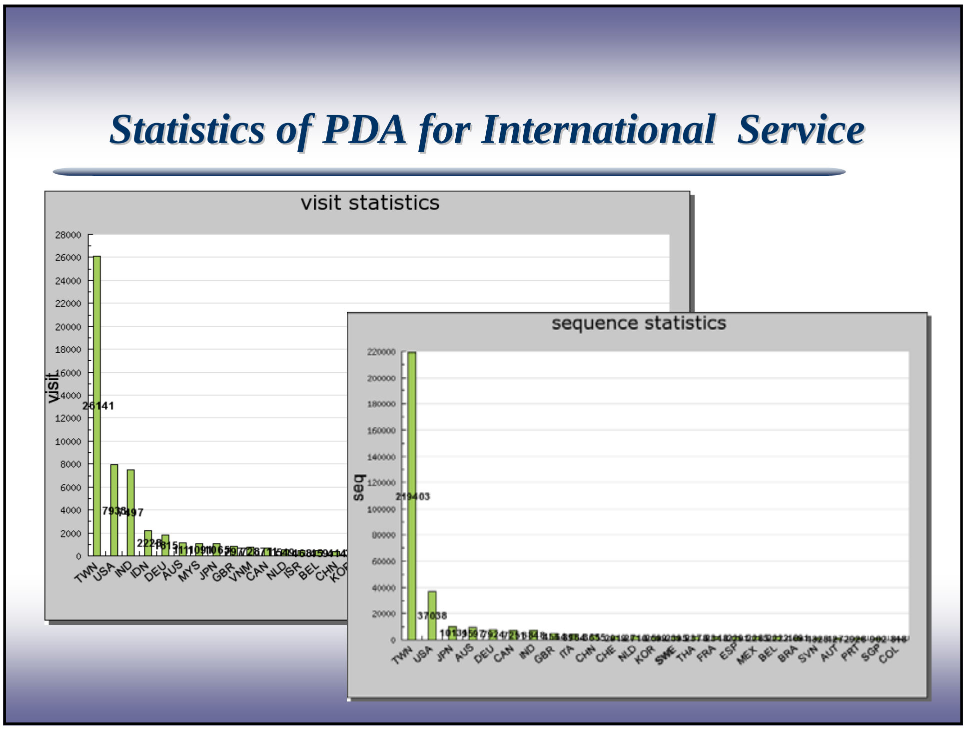
S.H. Chen[1], C.Y. Lin*, C.S. Cho, C.Z. Lo and C.A. Hsiung

Division of Biostatistics and Bioinformatics, National Health Research Institutes, #128, Yen-Chiu-Yuan Rd Sec. 2, Taipei 115, Taiwan and [1]Institute of Zoology, National Taiwan University, Taipei, Taiwan

Received February 15, 2003; Revised March 13, 2003; Accepted March 31, 2003



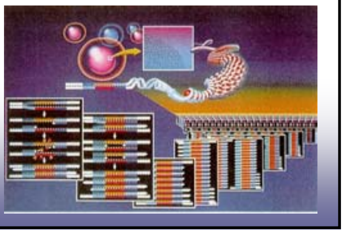
Accumulative Visitors and Sequences
during 2003. Jul ~ 2006. Oct



Statistics of PDA for International Service

Future Plan for PDA

- Based on the similar algorithm with principles from the wet lab, we will develop novel interfaces to design the primers/probes for Multiplex PCR, Real-time PCR, methylation PCR etc..
- Of course, all the applications will be online and accessed freely.



Pipeline for Multiple PCR Primer Design

Target sequences in fasta format

⬇

Select organism

⬇

Avoid over-presented region in target sequence

⬇

Based on thermodynamic theory and other critical issues, select primers near 3’

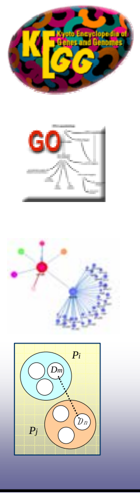
⬇

Ranking the candidate primers

⬇

Output on web in a session (in 60 sec)

Database of Protein Interactome, DPI



yeast ORF search

Search the Field : Full text

Enter the Keyword

You can input mutiple items. Each item is separeted by comma ','

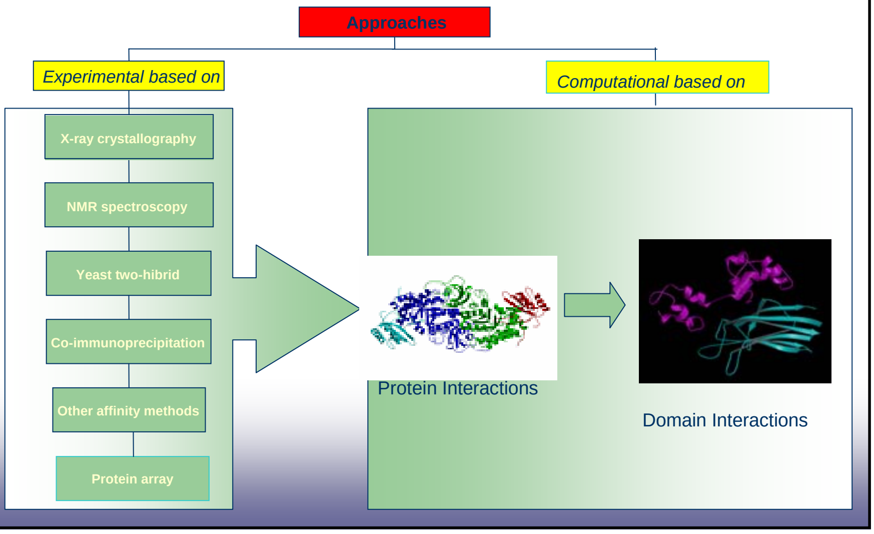
Order : default

select chromosome : All

Select Output Field

☑ Select all ☐ Unselect all
☑ orf ☑ SGD_ID ☑ Chromosome ☑ locus
☑ alias ☑ Description ☑ geneproduct ☑ phenotype
☑ Genebank_ID ☑ Protein_ID

New Approach for Interactomics



Approaches

Experimental based on

- X-ray crystallography
- NMR spectroscopy
- Yeast two-hibrid
- Co-immunoprecipitation
- Other affinity methods
- Protein array

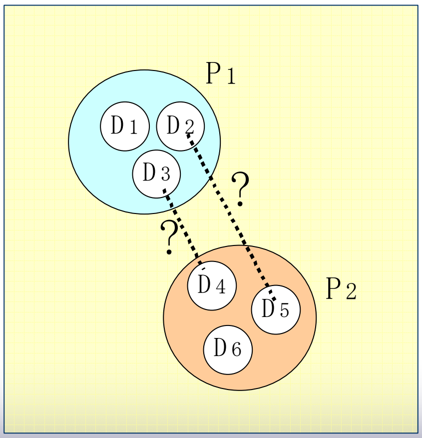
Computational based on

Protein Interactions

Domain Interactions

Inferred Protein Interactions

- Data from protein interactions
  - Wet lab produced
- Inferred protein interactions from domain interactions
  - Association method
  - EM(Expectation Maximization)- MLE value



Helicobacter pylori (幽門桿菌)-
Database of Protein Interactome (Bioinformatics, 2005)

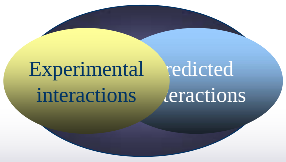
Letters to Nature

Nature 409, 211-215 (11 January 2001) | doi: 10.1038/35051615

The protein–protein interaction map of *Helicobacter pylori*

Jean-Christophe Rain[1], Luc Selig[1], Hilde De Reuse[2], Véronique Battaglia[1], Céline Reverdy[1], Stéphane Simon[1], Gerlinde Lenzen[1], Fabien Petel[1], Jérôme Wojcik[1], Vincent Schächter[1], Y. Chemama[1], Agnès Labigne[2] and Pierre Legrain[1]

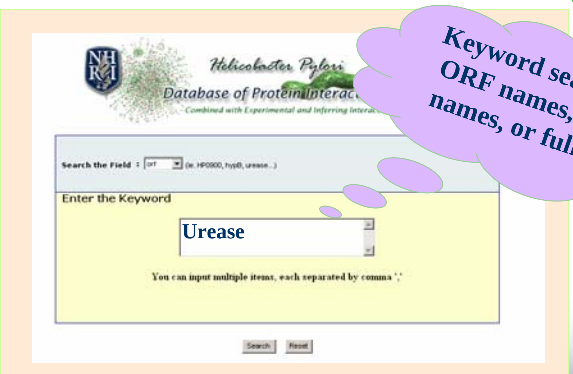
Over 1,200 interactions were identified between H. pylori (strain 26695) proteins, connecting 46.6% of the proteome.



The network of whole proteome

Service to be announced

Database of Protein Interactomes (DPI)



Search the Field : orf (ie. HP0900, hypB, urease...)

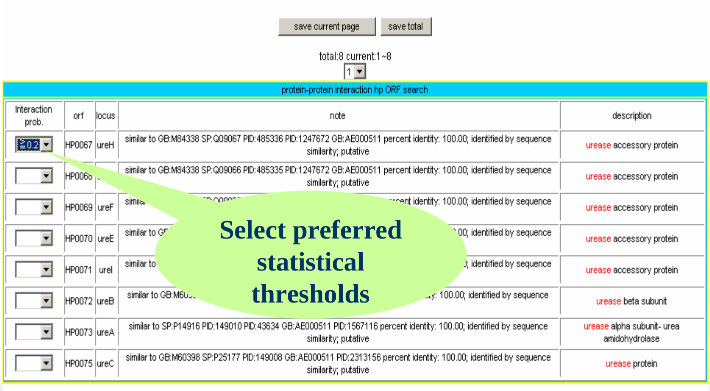
Enter the Keyword

Urease

You can input multiple items, each separated by comma ','

Keyword search by ORF names, locus names, or full text

Service to be announced (conti)



total:8 current:1~8

1

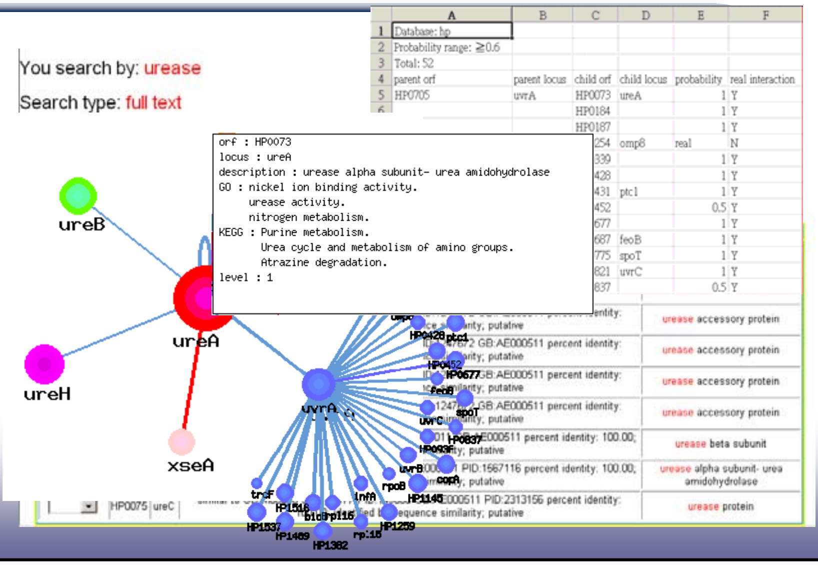
protein-protein interaction hp ORF search

Interaction prob.	orf	locus	note	description
≧0.2	HP0067	ureH	similar to GB:M84338 SP:Q09067 PID:485336 PID:1247672 GB:AE000511 percent identity: 100.00; identified by sequence similarity; putative	urease accessory protein
	HP0068	[illegible]	similar to GB:M84338 SP:Q09066 PID:485335 PID:1247672 GB:AE000511 percent identity: 100.00; identified by sequence similarity; putative	urease accessory protein
	HP0069	ureF	simila [illegible] rcent identity: 100.00; identified by sequence [illegible]	urease accessory protein
	HP0070	ureE	similar to G [illegible] 00; identified by sequence [illegible]	urease accessory protein
	HP0071	ureI	similar to [illegible]; identified by sequence [illegible]	urease accessory protein
	HP0072	ureB	similar to GB:M6 [illegible] y: 100.00; identified by sequence [illegible]	urease beta subunit
	HP0073	ureA	similar to SP:P14916 PID:149010 PID:43634 GB:AE000511 PID:1567116 percent identity: 100.00; identified by sequence similarity; putative	urease alpha subunit- urea amidohydrolase
	HP0075	ureC	similar to GB:M60398 SP:P25177 PID:149008 GB:AE000511 PID:2313156 percent identity: 100.00; identified by sequence similarity; putative	urease protein

Select preferred statistical thresholds

Report page of Database of Protein Interactomes

Search Result of hp-DPI



You search by: urease

Search type: full text

1 Database: hp

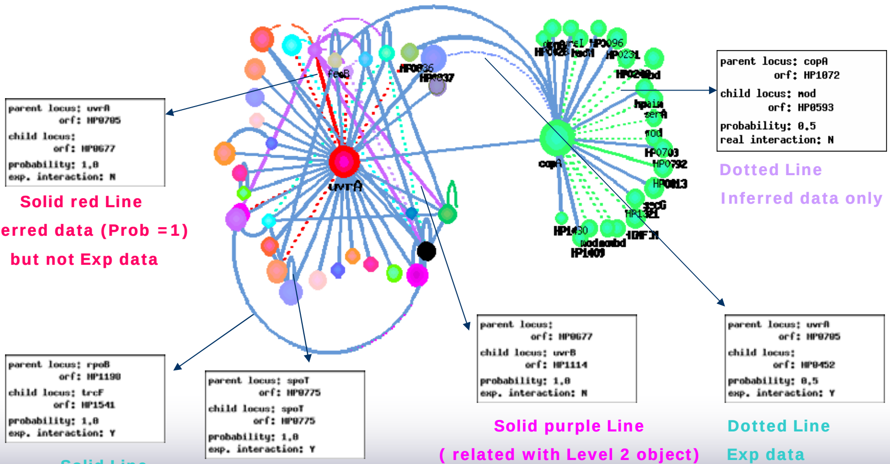
2 Probability range: ≧0.6

3 Total: 52

	A	B	C	D	E	F
4	parent orf	parent locus	child orf	child locus	probability	real interaction
5	HP0705	uvrA	HP0073	ureA	1	Y
6			HP0184		1	Y
			HP0187		1	Y
			254	omp8	real	N
			339		1	Y
			428		1	Y
			431	ptc1	1	Y
			452		0.5	Y
			677		1	Y
			687	feoB	1	Y
			775	spoT	1	Y
			821	uvrC	1	Y
			837		0.5	Y

```
orf : HP0073
locus : ureA
description : urease alpha subunit- urea amidohydrolase
GO : nickel ion binding activity.
     urease activity.
     nitrogen metabolism.
KEGG : Purine metabolism.
       Urea cycle and metabolism of amino groups.
       Atrazine degradation.
level : 1
```

Edges Patterns for Interaction



```
parent locus: uvrA
        orf: HP0705
child locus:
        orf: HP0677
probability: 1.0
exp. interaction: N
```

Solid red Line

Inferred data (Prob =1)

but not Exp data

```
parent locus: rpoB
        orf: HP1198
child locus: trcF
        orf: HP1541
probability: 1.0
exp. interaction: Y
```

Solid Line

Exp. Data and inferred Prob =1

```
parent locus: spoT
        orf: HP0775
child locus: spoT
        orf: HP0775
probability: 1.0
exp. interaction: Y
```

Solid curve -Self Interaction

Exp. Data and inferred Prob =1

```
parent locus:
        orf: HP0677
child locus: uvrB
        orf: HP1114
probability: 1.0
exp. interaction: N
```

Solid purple Line

(related with Level 2 object)

Inferred data (Prob =1)

but not Exp data

```
parent locus: copA
        orf: HP1072
child locus: mod
        orf: HP0593
probability: 0.5
real interaction: N
```

Dotted Line

Inferred data only

```
parent locus: uvrA
        orf: HP0705
child locus:
        orf: HP0452
probability: 0.5
exp. interaction: Y
```

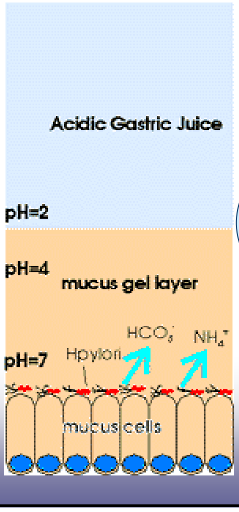
Dotted Line

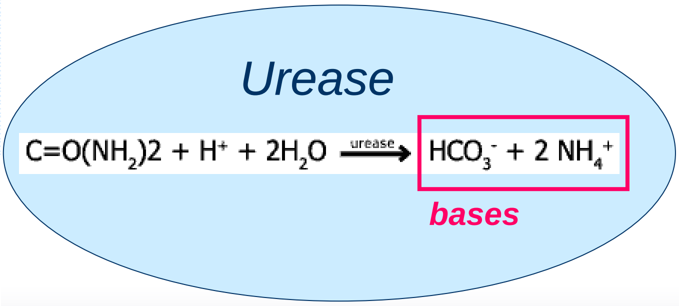
Exp data

but Inferred data

With Prob<1

Discover New Research Targets with hp-DPI

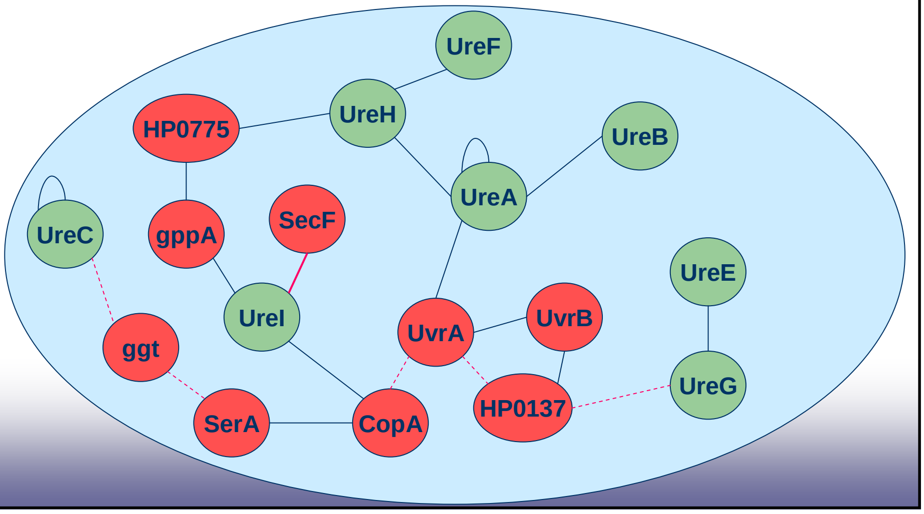



Urease

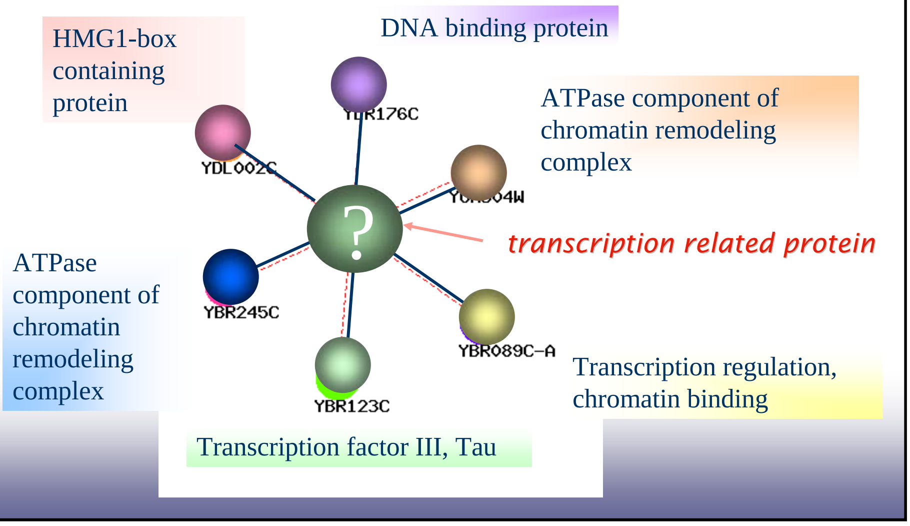
$$C{=}O(NH_2)2 + H^+ + 2H_2O \xrightarrow{urease} \boxed{HCO_3^- + 2\ NH_4^+}$$

bases

Network of Urease Complex



Annotated Protein Function



HMG1-box containing protein

DNA binding protein

ATPase component of chromatin remodeling complex

transcription related protein

ATPase component of chromatin remodeling complex

Transcription regulation, chromatin binding

Transcription factor III, Tau

---

hp-DPI (http://dpi.nhri.org.tw/hp/)

Institution: National Health Research Intitutes Sign In as Personal Subscriber

Bioinformatics, doi:10.1093/bioinformatics/bti101
Bioinformatics © Oxford University Press 2004; all rights reserved
Received June 24, 2004
Revised September 17, 2004
Accepted October 17, 2004

hp-DPI: *Helicobacter pylori* database of protein interactomes- embracing experimental and inferred interactions

Chung-Yen Lin [1*], Chia-Ling Chen [1], Chi-Shiang Cho [1], Li-Ming Wang [1], Chia-Ming Chang [1], Pao-Yang Chen [1], Chen-Zen Lo [1], and Chao A. Hsiungs [1]

[1] Division of Biostatistics and Bioinformatics, National Health Research Institutes. #128, Sec. 2 Yaun-Chio-Yan Rd. Taipei 115, Taiwan

[*] To whom correspondence should be addressed.
Chung-Yen Lin, E-mail: cylin@nhri.org.tw

hp-DPI Selected into
2006 The Molecular Biology Database Collection by NAR

OXFORD JOURNALS | CONTACT US | MY BASKET | MY ACCOUNT

Nucleic Acids Research

ABOUT THIS JOURNAL | CONTACT THIS JOURNAL | SUBSCRIPTIONS | CURRENT ISSUE | ARCHIVE | SEARCH

Oxford Journals › Life Sciences › Nucleic Acids Research › Database Summary Paper 664

« PREVIOUS NEXT »

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

hp-DPI

NAR Molecular Biology Database Collection entry number 664

http://dpi.nhri.org.tw/hp/

Database Description

Database of protein interactions in *Helicobacter pylori*

Category: Metabolic and Signaling Pathways
Subcategory: Intermolecular interactions and signaling pathways

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

Oxford University Press is not responsible for the content of external internet sites

Online ISSN 1362-4962 - Print ISSN 0305-1048

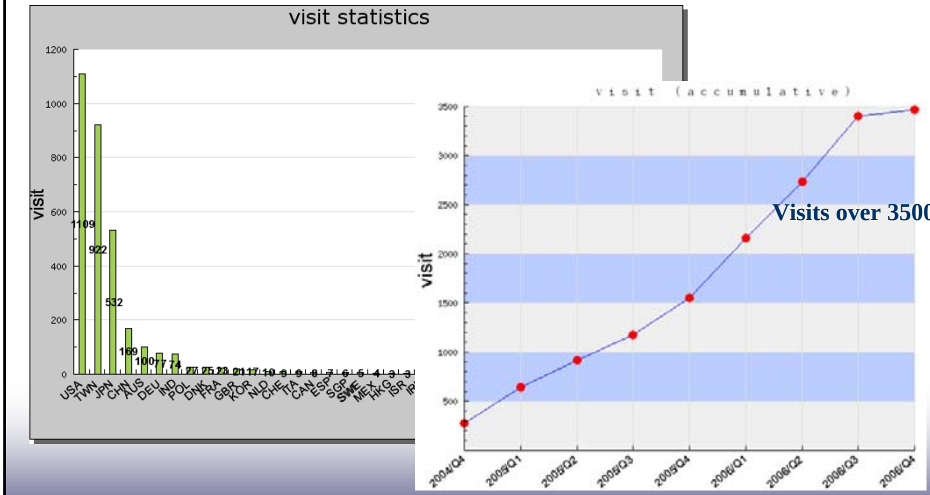
Copyright © 2006 Oxford Journals

OXFORD JOURNALS
OXFORD UNIVERSITY PRESS

Site Map | Privacy Policy | Frequently Asked Questions

Other Oxford University Press sites:
Oxford University Press

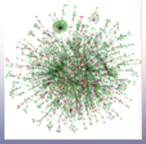
Visit Statistics for hp-DPI
from 2004/11/22 ~ 2006/10/11



Visits over 3500

Features in hp-DPI (BMC Bioinformatics, 2006, In press)

- Functions of hypothesized proteins can be speculated through such a network approach as well
- This system not only generates inferring interactions but also functions as an examining tool to point out possible false positive results derived from two-hybrid experiments.
- With LAMP, this system is able to provide high-quality and speedy services with free cost.



Fly Database of Protein Interactomes

yeast ORF search

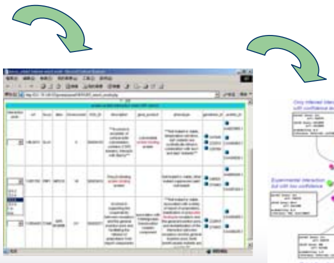
Search the Field : full_text

Enter the Keyword

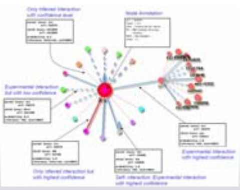
protein binding

You can input mutiple items. Each item is separeted by comma ','

Query in Full Text



Search Result & Statistical Estimation



Network Visualization & popup annotation

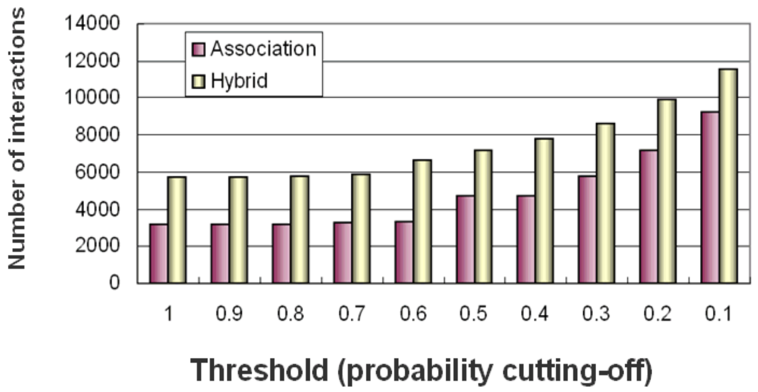
Available soon

Hybrid Model for Estimation

- **Maximized Likelihood Estimation (MLE)**
  - To incorporate the global consideration for whole interactimes and the experimental error ignored by local estimate, the likelihood function *L* is based on all potential protein interactions.

$$L = \prod \left(\Pr(o_{ij} = 1)\right)^{O_{ij}} \left(1 - \Pr(o_{ij} = 1)\right)^{1-O_{ij}}$$

The comparison of Association Measures and Hybrid Model -by the number of interactions predicted at each threshold



New Features in FlyDPI

Ping-Pong Search

Full-text Search

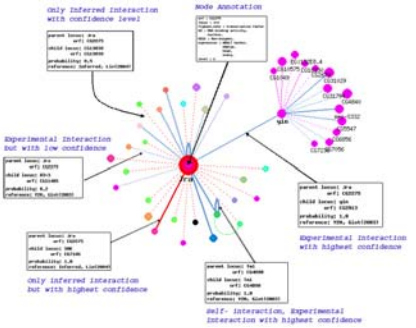
Chromosome Location

Gene Categories form GO

Spatiotemporal Scenarios

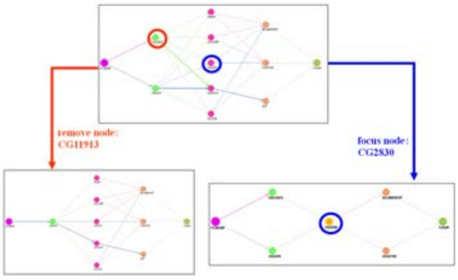
Search Results of FlyDPI

General Search



A snap of the experimental and inferred visualized interaction networks of *D. melanogaster* interactome under specific spatiotemporal scenarios.

Ping-Pong Search

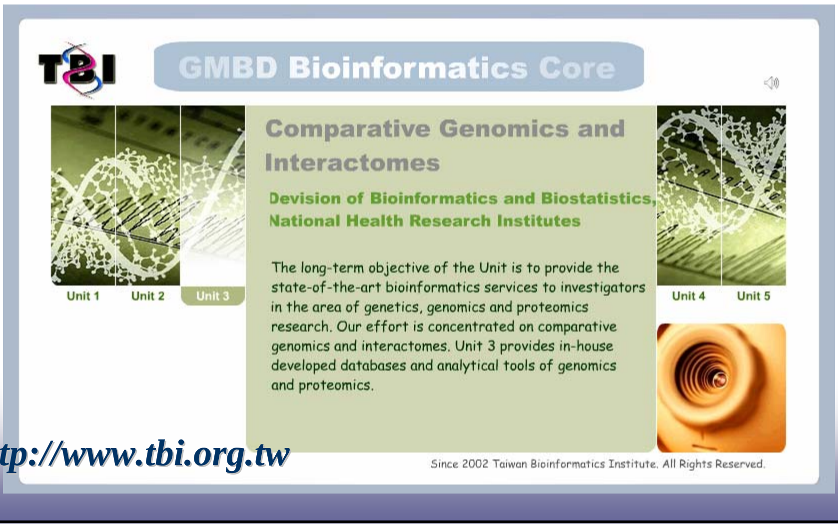


Map of proteins potentially involved in apoptosis generated by ping-pong search. By the click on the nodes or lines between two query proteins, the advanced option will remove the paths related or confine the paths with the selected nodes or lines

Acknowledgement

Chieh-Hwa Lin
Shu-Jun Hsu
Chia-Ling Chen
Fan-kai Lin
Li-Wei Lai
Chao A. Hsiung

Bioinformatics Core for Genomic Medicine and Biotechnology Development

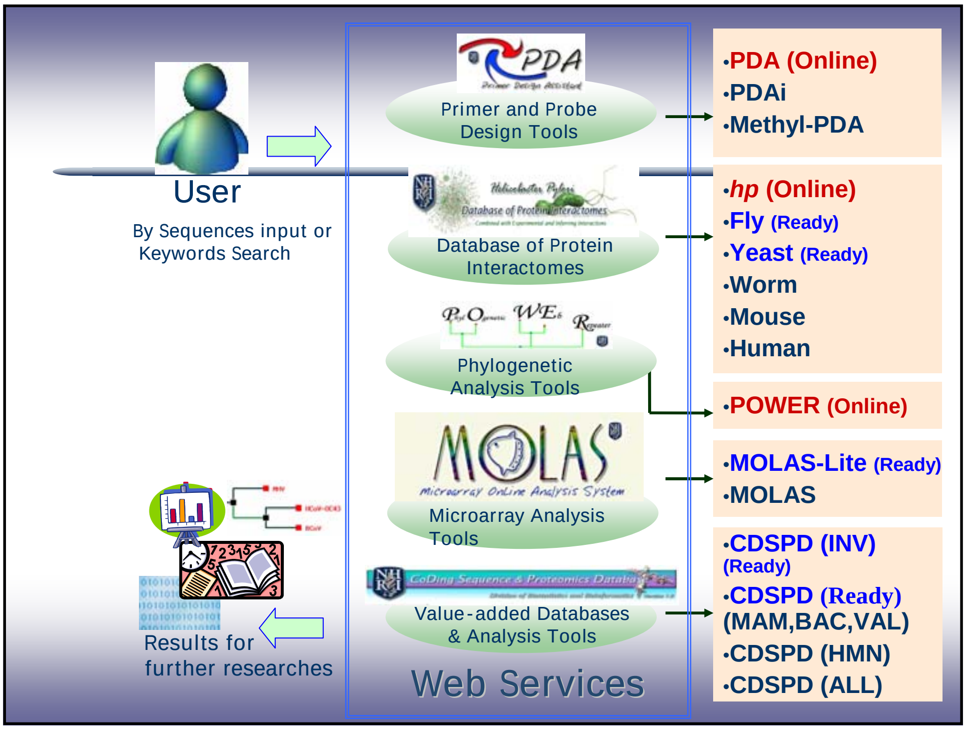


Comparative Genomics and Interactomes

Devision of Bioinformatics and Biostatistics, National Health Research Institutes

The long-term objective of the Unit is to provide the state-of-the-art bioinformatics services to investigators in the area of genetics, genomics and proteomics research. Our effort is concentrated on comparative genomics and interactomes. Unit 3 provides in-house developed databases and analytical tools of genomics and proteomics.

http://www.tbi.org.tw



User

By Sequences input or Keywords Search

Primer and Probe Design Tools

Database of Protein Interactomes

Phylogenetic Analysis Tools

Microarray Analysis Tools

Value-added Databases & Analysis Tools

Results for further researches

Web Services

- **PDA (Online)**
- **PDAi**
- **Methyl-PDA**

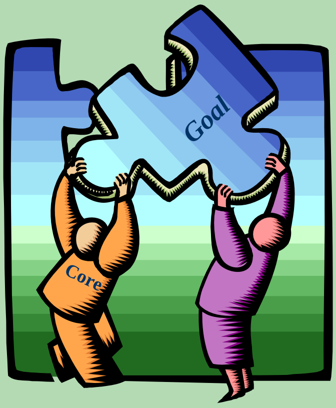
- ***hp* (Online)**
- **Fly (Ready)**
- **Yeast (Ready)**
- **Worm**
- **Mouse**
- **Human**

- **POWER (Online)**

- **MOLAS-Lite (Ready)**
- **MOLAS**

- **CDSPD (INV) (Ready)**
- **CDSPD (Ready) (MAM,BAC,VAL)**
- **CDSPD (HMN)**
- **CDSPD (ALL)**

Thanks for Your Attention!
And Time for Demonstration



國家衛生研究院
National Health Research Institutes
Division of Biostatistics and Bioinformatics

中央研究院
資訊科學研究所
Institute of Information Science
Academia Sinica
Systems Biology and Network Biology Laboratory