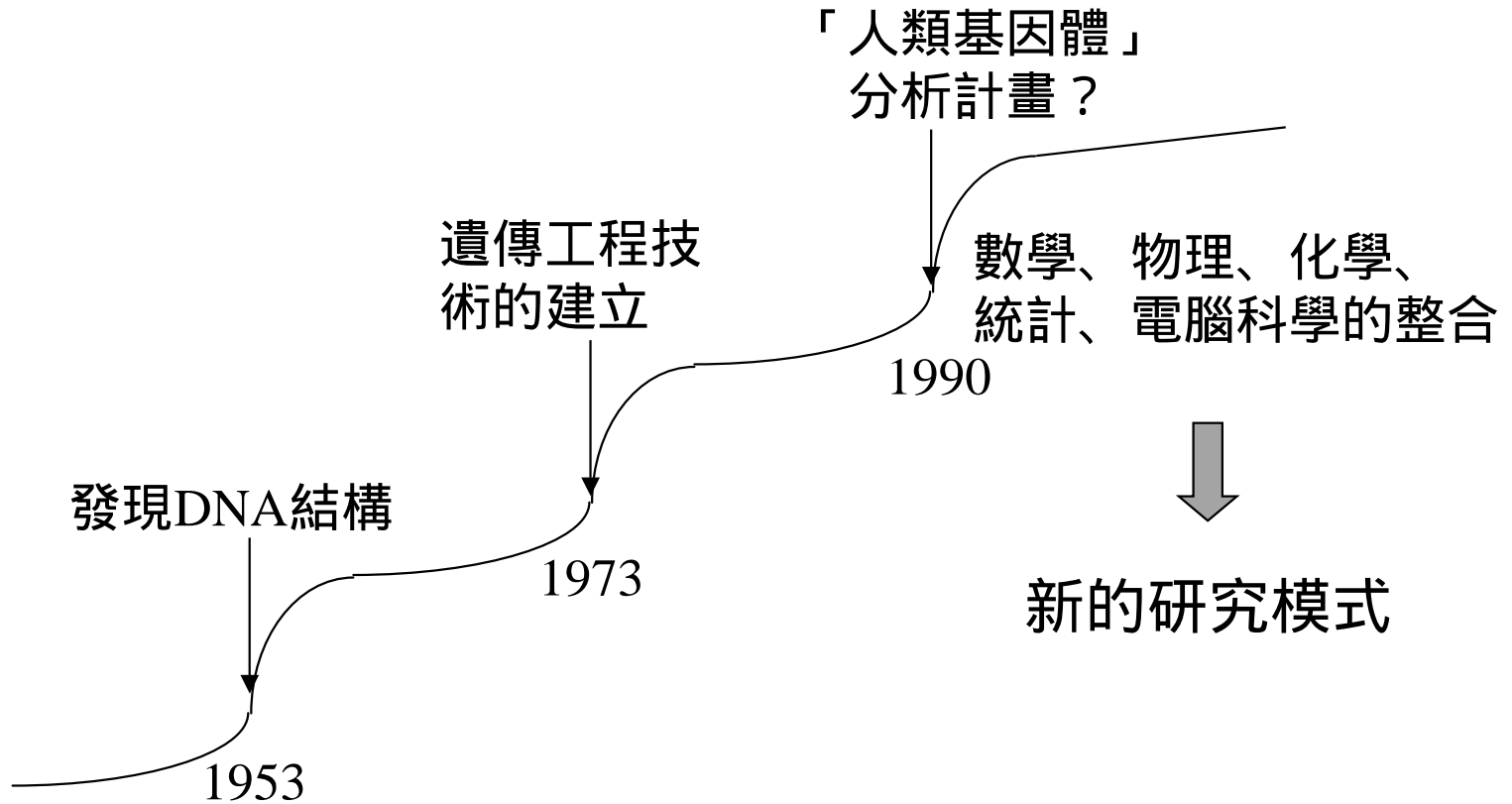


蛋白質之生物資訊分析簡介

陽明生化所、生物資訊學程、
生物資訊中心 楊永正

觀察歷史有助於了解時代趨勢



新的研究模式

鉅量分析 (high throughput analysis)



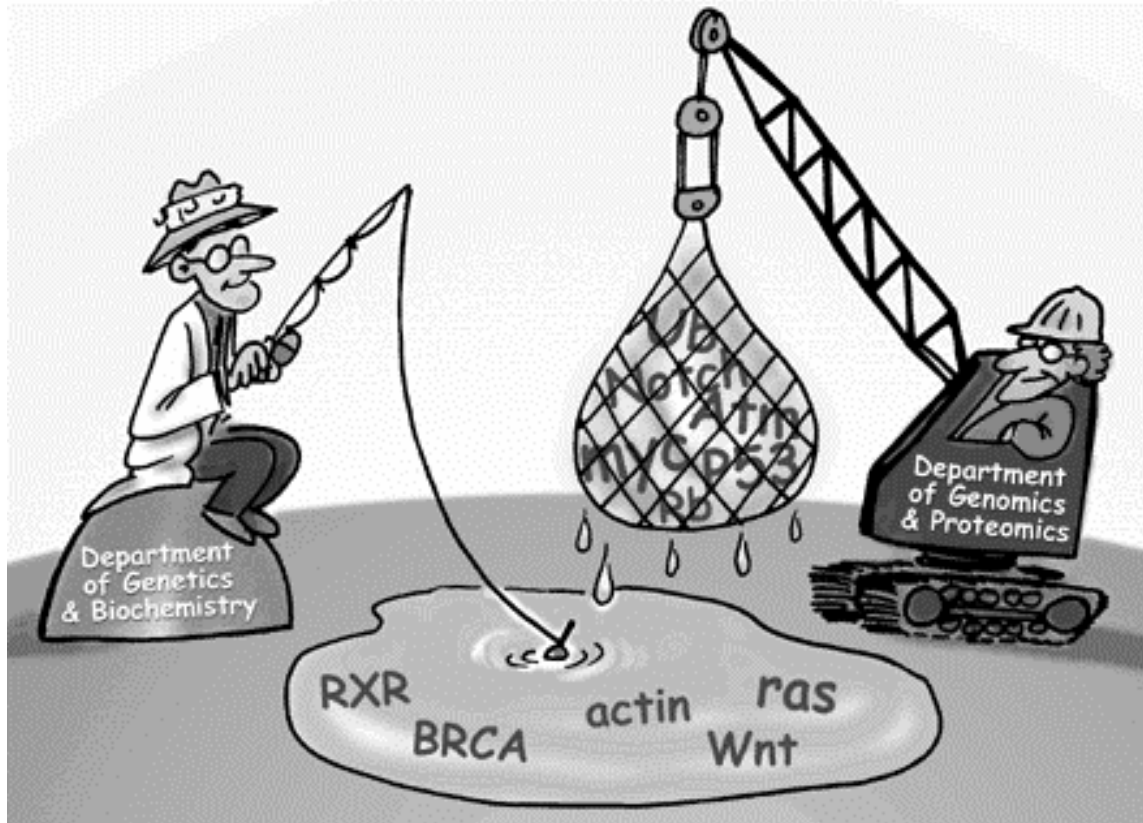
得到所有的資訊



宏觀分析 (global analysis)

(與傳統上“研究少數蛋白質”與
“由假說設計實驗”的方法不同)

傳統分析 vs 鉅量分析

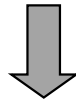


<http://www.sciencemag.org/cgi/content/full/291/5507/1221/F1>

YM-Biochem

生物學研究的最大難題

尋找基因的功能



找到更多的基因能解決問題嗎？

DNA 語言為何難懂？

5 10 15 20 25 30 35
GARBA GEYOU THROW OUTJU NKYOU KEEP HERE F
40 45 50 55 60 65 70
ORE90 % OF THE GENOME IS JUNK
MEISJ UNKSB RENNE R

Garbage you throw out, junk you keep. Therefore, 90% of the genome is junk. - S. Brenner.

語言的三大要素： 字,詞與文法

**Garbage you throw out, junk you keep. Therefore,
90% of the genome is junk. - S. Brenner.**



A->B, B->C, C->D,
D->E, E->F, --- *etc.*

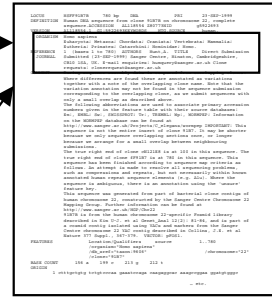
Hbscbhf zpv uispv pvu, **kvof** zpv lffq. Uifsfgpsf,
01^ fg uif hfopnf jt **kvof**. =T. Csfoofs.

語言學的類比 - 字與詞

Junk

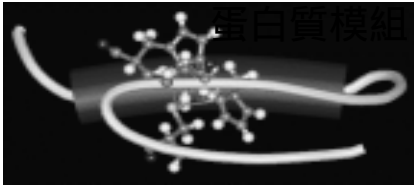
Garbage you throw out, junk you keep. Therefore, 90% of the genome is junk.

- S. Brenner.

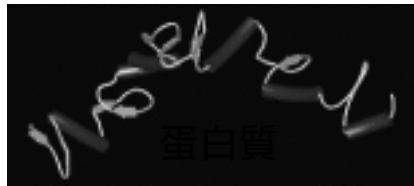


生命之書

字

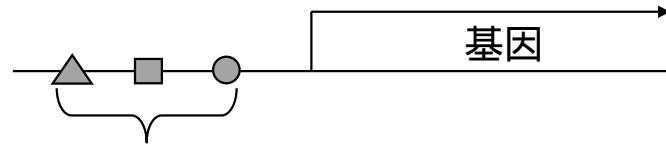


詞



獨立折疊單元

轉錄因子接合位置



調控區

序列決定因子

蛋白質之生物資訊分析

(<http://ymbc.ym.edu.tw/proteome/>)

- 二維電泳相關分析
- 質譜儀相關分析
- 蛋白質資料庫
- 蛋白質分析
- 蛋白質交互作用
- 蛋白質結構
- 反應路徑

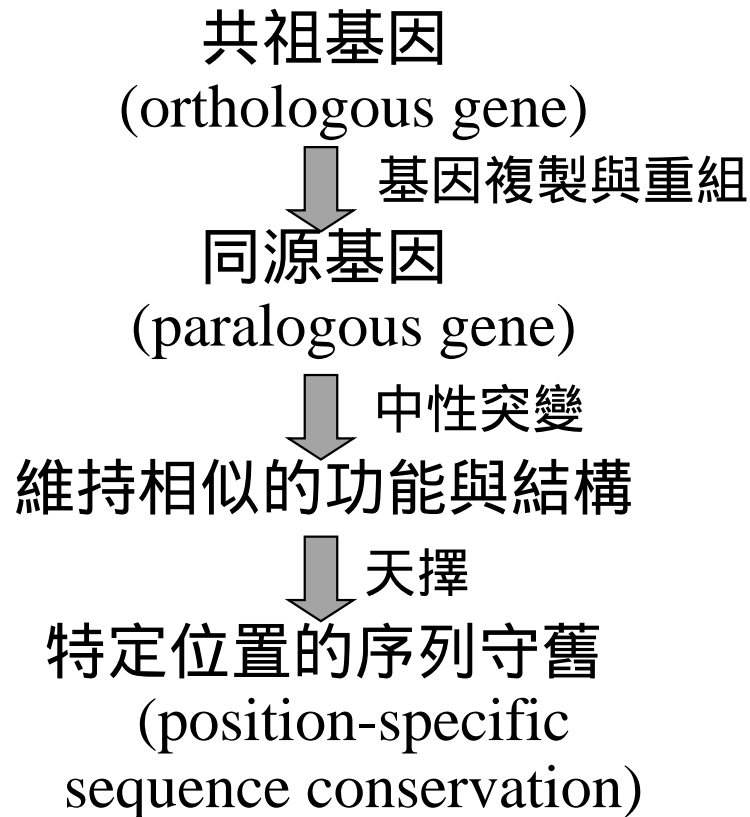
生物資訊應用的兩大要件

生物資訊在哪裡？
如何選用適當的工具？

蛋白質資料庫

- Protein information resource (PIR)
 - Protein sequence database (PSD)
 - Annotation and Similarity database (ASDB)
 - NRL3D
 - RESID
 - ALIGN
 - SwissProt & Trembl
 - GenPept (NCBI)
 - OWL
 - non-redundant composite of 4 publicly-available primary sources: *SWISS-PROT*, *PIR (1-3)*, *GenBank* (translation) and *NRL-3D*.
- (<http://www.bioinf.man.ac.uk/dbbrowser/OWL/>)

蛋白質模組的形成



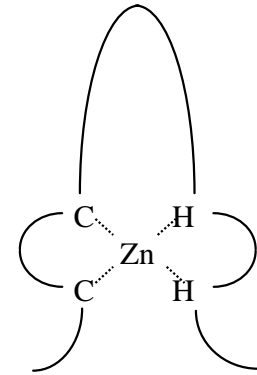
蛋白質模組的表示方法

共識序列

位置加權矩陣

隱藏式馬可夫模型

共識序列 – 範例



Patterns of metal-binding domains described in Berg, J.M. (1986)
Potential metal-binding domains in nucleic acid binding proteins.
Science 232, 485-487

Name	Offset	Pattern ..
TFIIIA	1	CX{2,5}CX{12,12}HX{2,3}H

(GCG的樣式表示方式)

共識序列資料庫

名稱: Prosite

負責人: Amos Bairoch

方法:由文獻中蒐集資料整理而成

樣式分析工具

<http://saw.ym.edu.tw/emboss/>

- Fuzzpro
 - Protein pattern search
- Patmatdb
 - Search a protein sequence database with a motif
- Patmatmotifs
 - Search a motif database with a protein sequence

在 EMBOSS 環境下寫模組樣式(pattern) 的規則

1. 胺基酸用一個字母的標準代碼(one-letter code) 表示。
 2. X 是代表任意的胺基酸。
 3. 胺基酸後若有小括號，則代表此胺基酸重覆的次數。一般而言，小括號中有兩個數字以逗點隔開，以表示一個範圍的起點與終點，例如 CX(2, 4)表示在 Cys 後有 2 至 4 個任意的胺基酸。若小括號中只有一個數字，則表示重覆的次數。
 4. 胺基酸若出現在中括號裡，如[ALT]則代表為Ala 或 Leu 或 Thr。
 5. 胺基酸若出現在大括號裡，如{AM}則代表為除了 Ala 及 Met 以外的胺基酸。
 6. 若想表示模組樣式出現在N- 或 C-terminal ，則分別以`<' 以及`>' 代表之。
- 以 [DE](2)HS{P}X(2)PX(2,4)C 為例，寫出的模組樣式表示在兩個 Asps 或 Glus 後是 His 以及 Ser，然後再接一個非 Pro 的胺基酸；一個 Pro 緊接在兩個任意胺基酸後；最後是在二到四個任意胺基酸後接著一個 Cys。

使用共識序列的問題

DEDEEE太嚴苛;

(E,D)E(E,D)(E,D)(E,D)(E,D)太鬆散,下列序列片段均有可能:

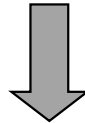
- DEDDDD
- EEDDDD
- DEEEEEE
- EEEEEEE
- ...等

序列的變異可微調巨分子間的交互作用

- 範例
 - 抗原與抗體間的交互作用
 - 轉錄因子與啟動子間的辨識
 - ... 等.
- ⇒如何解決共識序列太嚴苛或太鬆散的問題?

Profile的概念

Position dependent
sequence conservation



Position specific
scoring matrix
(PSSM)

Gribskov, M., Homyak, M., Edenfield, J., Eisenberg, D. (1988) Profile scanning for three-dimensional structural patterns in protein sequences. *Comput Appl Biosci.* 4, 61-66.

利用PSSM尋找序列模組

	C	T	A	T	A	A	T	C	Total Score
A	-38	19	1	12	10	-48			
C	-15	-38	-8	-10	-3	-32			
G	-13	-48	-6	-7	-10	-48			
T	17	-32	8	-9	-6	19			
Score	-15	-32	1	-9	10	-48			-93



	C	T	A	T	A	A	T	C	Total Score
A		-38	19	1	12	10	-48		
C		-15	-38	-8	-10	-3	-32		
G		-13	-48	-6	-7	-10	-48		
T		17	-32	8	-9	-6	19		
Score		17	19	8	12	10	19		85

Shift PSSM along the seq and sum up the score

最高分處即為模組 所在的位置

	C	T	A	T	A	A	T	C	Total Score
A		-38	19	1	12	10	-48		
C		-15	-38	-8	-10	-3	-32		
G		-13	-48	-6	-7	-10	-48		
T		17	-32	8	-9	-6	19		
Score		17	19	8	12	10	19		85



	C	T	A	T	A	A	T	C	Total Score
A		-38	19	1	12	10	-48		
C		-15	-38	-8	-10	-3	-32		
G		-13	-48	-6	-7	-10	-48		
T		17	-38	8	-9	-6	19		
Score		-38	-38	1	12	-6	-32		-101

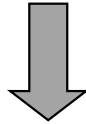
Profile分析工具

<http://saw.ym.edu.tw/emboss/>

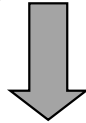
- Profit
 - Scan a sequence or database with a matrix or profile
- Prophecy
 - Creates matrices/profiles from multiple alignments
- Prophet
 - Gapped alignment for profiles

利用PSSM概念建構的模組資料庫: Blocks & Prints

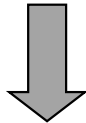
Get related sequences based on Prosite



Multiple sequence alignment



Position dependent scoring matrix

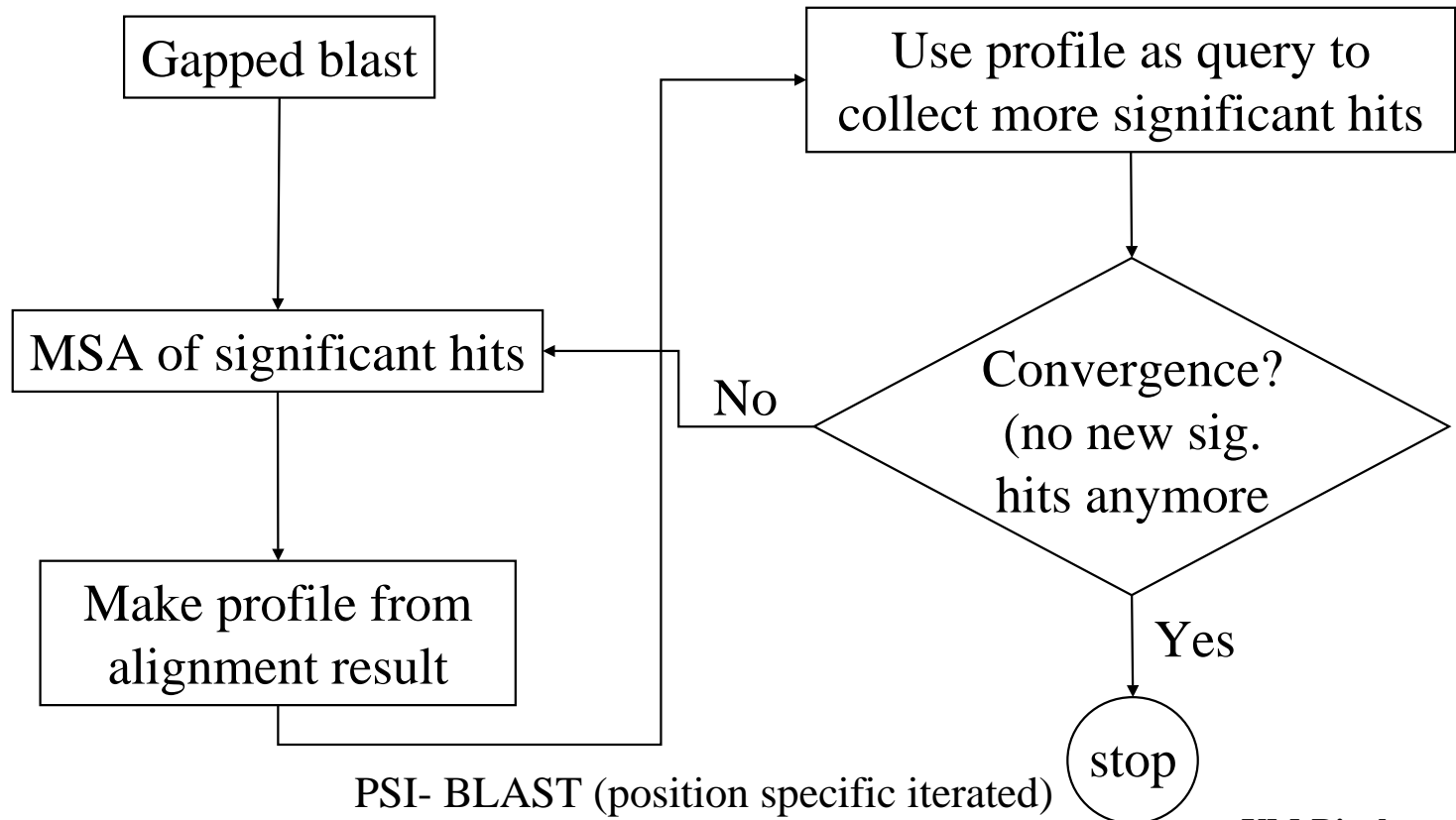


Take the most conserved region as blocks
(automatic) or prints (manual)

使用blocks或prints尋找蛋白質家族

All the blocks or prints that characterize a domain should all exist

利用PSSM以遞迴的方式 搜尋序列 資料庫,可尋找未發現的模組



PSI-blast的應用

A more sensitive search method

Discover novel motifs

利用PSI-blast概念建構的 模組資料庫: ProDom

- Created by exhaustive PSI-blast analysis for those protein families defined in Prosite
- Cluster sequences and present them in a tree form
- Connect to structure-prediction servers, ... *etc.*

Pattern-Hit-Induced blast (PHI-blast)

與PSI-blast相似，
但輸入是“樣式”而非序列

表示樣式(pattern)的語法

- [LFYT] = L or F or Y or T
- x(5) = xxxxx (x = any residues)
- x(2,4) = xx or xxx or xxxx
- Example:
 - ID ER_TARGET; PATTERN. PA [KRHQSA]-[DENQ]-E-L>. HI (19 22) HI (201 204)
- * <http://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html>

Profile的問題

- 不同位置間有空隙或相關, ...

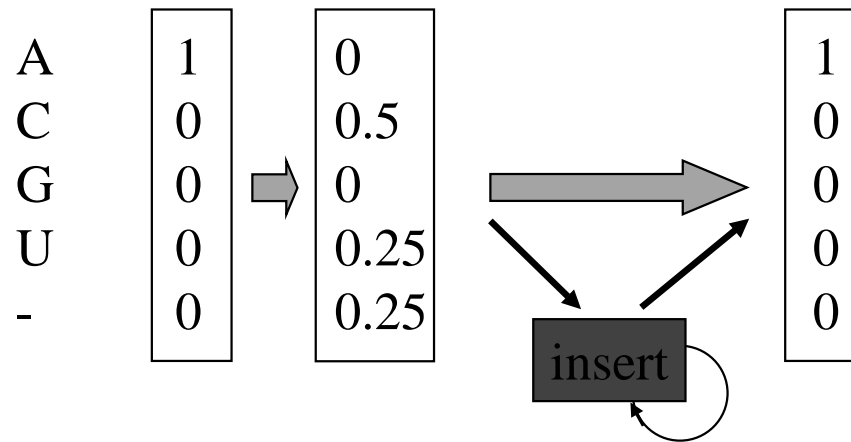
rat	A	C	G	G	A
ecoli	A	C	-	-	A
cow	A	-	-	-	A
corn	A	U	-	-	A

A	1	0	0	0	1
C	0	0.5	0	0	0
G	0	0	0.25	0.25	0
U	0	0.25	0	0	0
-	0	0.25	0.75	0.75	0

* This example was taken from Richard Hughey's handouts

解決方案 – 利用 “state” 與 “state”間之轉移機率

rat	A	C	G	G	A
ecoli	A	C	-	-	A
cow	A	-	-	-	A
corn	A	U	-	-	A



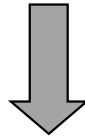
隱藏式馬可夫模型

Hidden Markov Model (HMM)

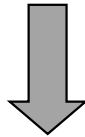
Patterns , profiles, ... *etc.*
可被視為 HMM 中的特例

利用HMM概念建構的 模組資料庫: pfam

Get related sequences based on Prosite



Multiple sequence alignment



Establish HMM model from
sequence alignment

模組的種類

- 酵素模組(接合與催化)
 - Relatively more conserved as they are directly related to function. They can be easily found by automatic procedures
- 調控模組(巨分子辨識)
 - There are lots of sequence variations, because these domains may not have a strong selection pressure. As a result, it is difficult to find by automatic method.

SMART

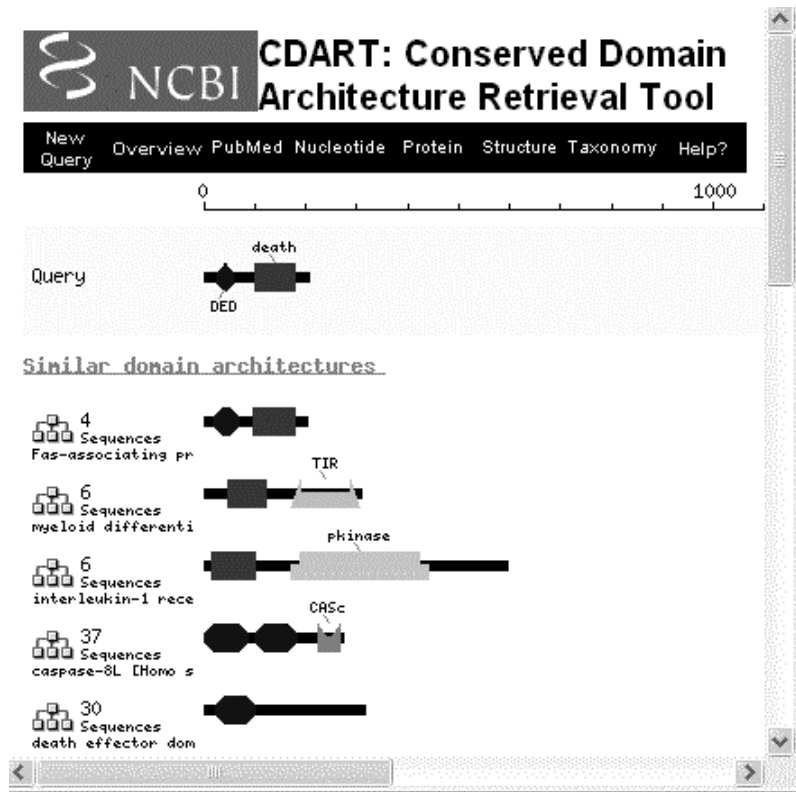
- 目標
 - to compile regulatory domains, such as those in the signal pathway, ... *etc.*
- 方法
 - PSI-blast analysis and manual curation
 - Refer to tertiary structure in defining domains

尋找序列中“所有的”模組

- 模組資料庫: CDD (conserve domain database)
 - Pfam: <http://pfam.wustl.edu/>
 - smart (simple modular architecture research tool): <http://smart.embl-heidelberg.de/>
- 搜尋工具
 - CD search: use RPS-blast (reverse position specific blast) to search for domains in a given sequence

搜尋蛋白質間相似的模組組合模式

- CDART: Conserved Domain Architecture Retrieval Tool
 - Use data from precomputed RPS-blast analysis
- SMART: Simple modular architecture research tool
 - Use precomputed domain search results



<http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps>

觀念整理

- 序列分析的三種模式
- 模組資料庫的種類
- 使用模組資料庫的兩種方法
- 與PSSM-相關的blast分析
- 蛋白體資訊分析的進一步資訊
 - <http://ymbc.ym.edu.tw/proteome/>
- 討論諮詢區
 - <http://binfo.ym.edu.tw/idg/>

序列分析的三種模式

字串搜尋 (String search)

樣式搜尋 (Pattern search)

相似性搜尋 (Similarity search)

模組資料庫的種類

- 利用文獻人工建構
 - Prosite
- 利用PSSM概念建構
 - Blocks (automatic)
 - Prints (manual)
- 利用PSI-blast概念建構
 - Prodom
- 利用HMM概念建構
 - Pfam
- 考慮調控模組
 - Smart
- 綜合
 - CDD = pfam + smart
 - InterPro = prosite + pfam + prints + prodom

使用模組資料庫的兩種方法

- 以單一序列搜尋模組資料庫
(尋找指定序列中的所有蛋白質模組)
- 以單一模組搜尋序列資料庫
(尋找所有含有指定模組的蛋白質序列)

與PSSM-相關的blast分析

- 尋找新的蛋白質模組
 - 以PSI-blast搜尋序列資料庫
- 較靈敏的資料庫搜尋方法
 - 以PSI-blast, PHI-blast搜尋序列資料庫
- 尋找指定序列中的所有蛋白質模組
 - 以RPS-blast搜尋CDD資料庫
- 搜尋蛋白質間相似的模組組合模式
 - CDART (Conserved Domain Architecture Retrieval Tool)
 - Smart (Simple Modular Architecture Research Tool)



Home

Mission statements

This web site is a joint effort of Yang-Ming Bioinformatics Research Center (YMBC) & Proteome Research Center (YMPC). YMPC tries to collect proteome information and YMBC tries to compare, analyze, and integrate the information from genomics, gene expression, and proteome. "Proteome informatics" is only part of the information system of YMBC, which focus on the proteome analysis. The goals of this site are

1. A portal site for proteome information
2. A site to host tools developed at YMBC
3. A site to host proteome information collected by YMU
4. A site to host proteome information generated at YMU

Hopefully, this site will promote proteome-related research at YMU.

Definition for Proteome (from Proteome Society)

1. The total protein complement of a genome.
2. Complete set of proteins expressed by a cell, tissue, or organism

[Home](#)

[Application examples](#)

[2D gel-related](#)

[MS-related](#)

[Protein db](#)

[Protein analysis](#)

[Protein interaction](#)

[Protein structure](#)

[Pathways](#)

討論諮詢區 <http://binfo.ym.edu.tw/idg/>

BioInfoFab



嗨, guest

會員:

密碼:

會員服務

全文搜尋

即時新聞


站務公告

歡迎提供任何新聞、小道消息、謠言、趣聞等等

如果你有任何議題想要開版討論歡迎在此申請

煩請各位讀者點選上

最新消息

 **成人幹細胞在其他器官扮演的角色..轉載自華文生技網**
刊登人: [醫明生資](#) 上線時間: 2002/03/13 @09:41AM 點閱率: 208

根據新英格蘭醫學期刊 (New England Journal of Medicine) 報導, 常被移植用來重建癌症病人血球供應系統的骨髓幹細胞, 可以在身體其他部分不同的器官自行整合, 這項發現為組織再生和器官移植帶來新契機.....(轉載自華文生技網)

[詳細內容...](#)(討論篇數: 00)

賽亞基因科技舉行新址啟用典禮
刊登人: [醫明生資](#) 上線時間: 2002/03/04 @05:47PM 點閱率: 308




賽亞基因科技昨日(3/3)舉行新址啟用典禮, 中研院長李遠哲、教育部長黃榮村、台北縣長蘇貞昌、台南科學園區籌備處主任戴謙等應邀參加...

[詳細內容...](#)(討論篇數: 00)

以動物實驗研究食慾控制機制..轉載自華文生技網
刊登人: [醫明生資](#) 上線時間: 2002/02/26 @03:20PM 點閱率: 232

約翰霍普金斯大學 (Johns Hopkins University School of Medicine) 的研究人員日前發表了他們關於減重及降低食慾研究的實驗數據。報告中指出, 瘦老鼠很快就會對每日服用降低食慾的藥物產生適應性, 而胖老鼠不會有這樣的情形.....(轉載自華文生技網)

[詳細內容...](#)(討論篇數: 00)

- 常駐論壇**
- [303] 生物資訊在哪裡?
 - [194] 生物技術交流 
 - [167] 生資名詞知多少? 
 - [152] 資訊學園地
 - [151] 生資工具使用經驗交流
 - [129] 一千零一個為什麼?
 - [101] 演講及研討會 
 - [067] GCG套裝軟體
 - [061] EMBOSS套裝軟體
- [\[更多論壇 | 新增論壇\]](#)

- 網站特搜**
- [1] Bio-Mirror
 - [2] BioMedGate Portal
 - [3] Genome Biology
 - [3] Genome Research
 - [3] 台大電機 Maxwell BBS Linux 版精華區

The End