

以生物資訊工具確認轉錄因子及受其調控之目標基因

Shih-Chieh Lin

*The Institute of Basic Medical
Sciences of NCKU*

07/19/2012

Part I:

Basic Concept of Gene Regulation

Part II:

Prediction of Transcription Factor Binding Sites

Part III:

Whole Genome Search of Transcription Factor Regulated Candidate Genes by Profile Hidden Markov Model

Transcriptional level control

- **Differential gene transcription – the major mechanism of selective protein synthesis**
- **Governed by a large number of proteins known as *transcription factors***
- **Two functional classes of transcription factors**
 - **General transcription factors**
 - **Specific transcription factors**

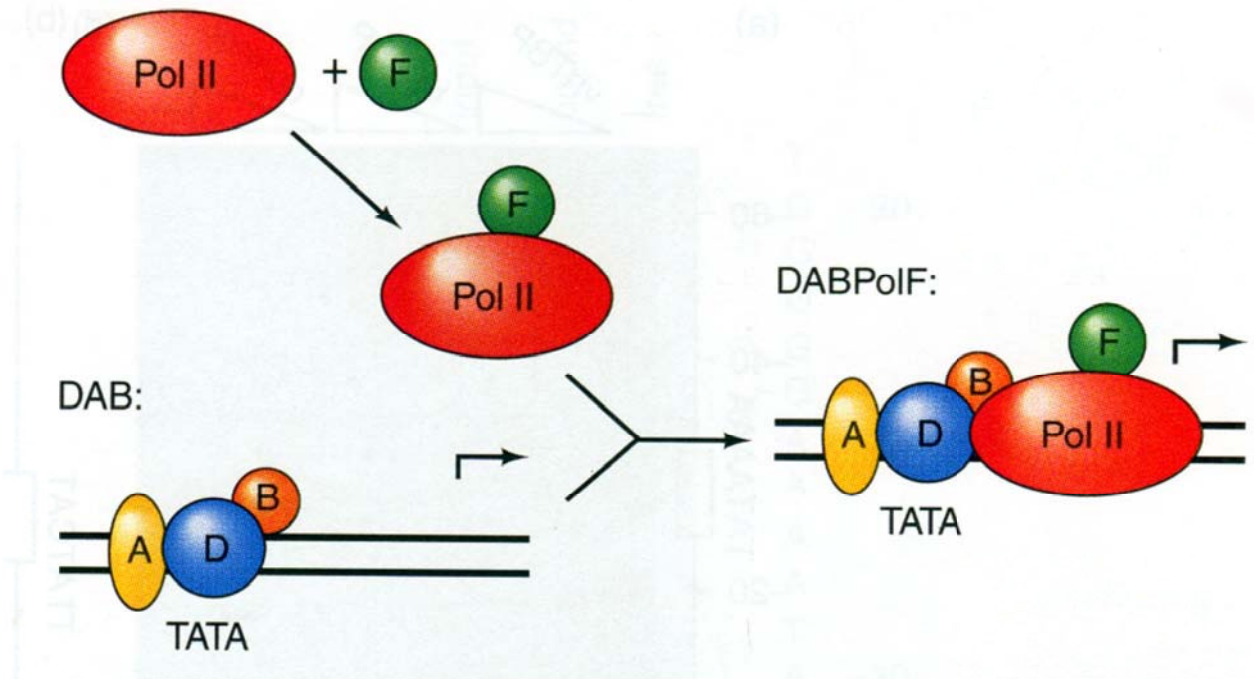
General Transcription Factors

Definition: combine with RNA polymerase to form a preinitiation complex that is competent to initiate transcription as soon as nucleotides are available

Class II Factors

TFIIA, TFIIB, TFIID,
TFIIE, TFIIF, TFIIH

TBP (TATA
box binding protein)

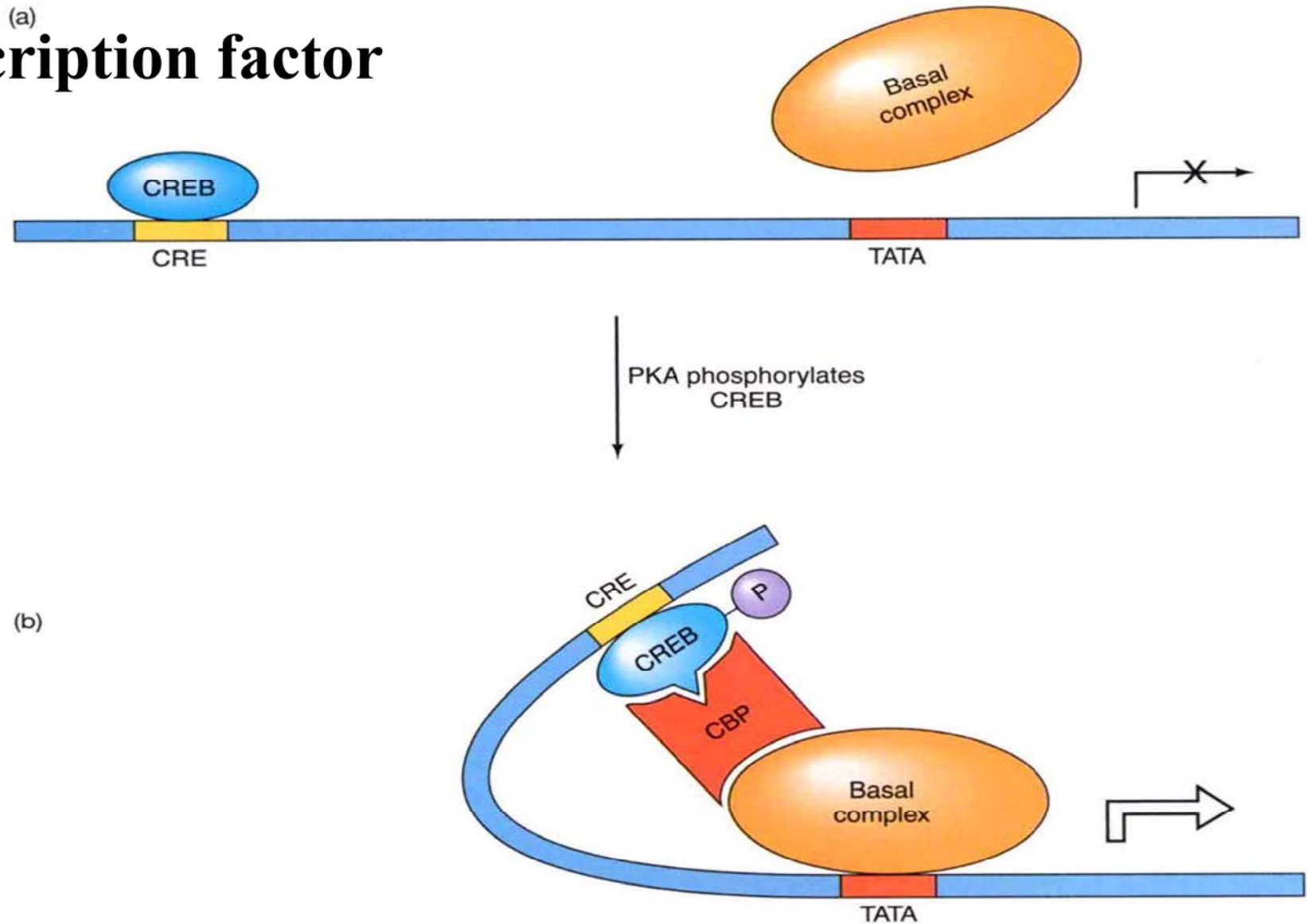


Specific transcription factors

- **A single gene controlled by many regulatory sites – bind different regulatory proteins**
- **A single regulatory protein may become attached to numerous sites on the genome**
- **Cells respond to environmental stimuli by synthesizing different transcription factors**
 - **Bind to different sites on DNA**

Control of Gene Expression

- **Activation of Transcription**
 - Transcription factor



Transcription Factors Binding Element

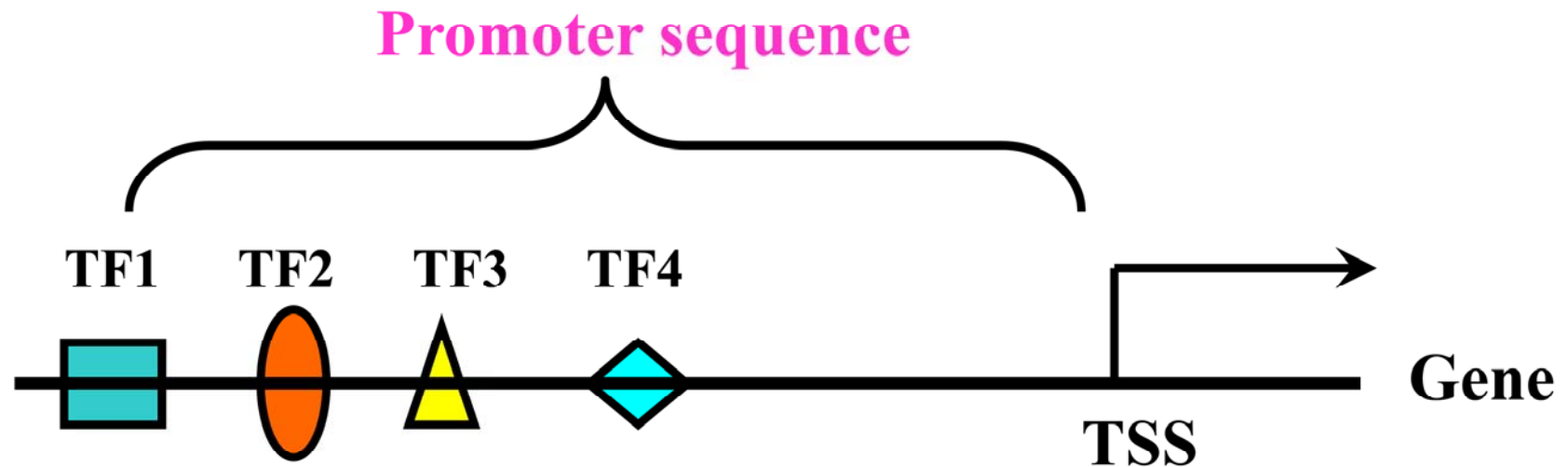
- **GRE**

5'-AGAACA_{nnn}TGTTCT-3'

3'-TCTTGT_{nnn}ACAAGA-5'

- **A palindrome**
- **Two-fold nature is important**
- **Pairs of GR polypeptides bind to DNA forming dimers**

Predict transcription factor binding element by bioinformatics tools



- **Signal Scan**

<http://www-bimas.cit.nih.gov/molbio/signal/>

- **TFSEARCH**

<http://www.cbrc.jp/research/db/TFSEARCH.html>

- **TESS**

<http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=SEA-FR-QueryS>

- **PROMO**

http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3

TSS: transcriptional start site

Part II:

Prediction of Transcription Factor Binding Sites

How to extract promoter sequence of gene ?

<http://www.ensembl.org/index.html>

Ensembl Genome Browser

 BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login · Register

Search: for
e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes ([Log in to customize this list](#))

 **Human**
GRCh37

 **Mouse**
NCBIM37

 **Zebrafish**
Zv9

All genomes

New to Ensembl?

Did you know you can:

-  [Learn how to use Ensembl](#)
with our video tutorials and walk-throughs
-  [Add custom tracks](#)
using our new Control Panel
-  [Upload and analyse your data](#)
and save it to your Ensembl account
-  [Search for a DNA or protein sequence](#)
using BLAST or BLAT
-  [Fetch only the data you want](#)
from our public database, using the Perl API
-  [Download our databases via FTP](#)
in FASTA, MySQL and other formats
-  [Mine Ensembl with BioMart](#)
and export sequences or tables in text, html, or Excel format

Still got questions? Try our [FAQs](#) or [glossary](#)

Did you know...?

 **FAQs**

Questions? ☐ Check out our [FAQs](#).

Search Ensembl

Search: for

e.g. human gene BRCA2 or rat X:100000..200000 or insulin

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes ([Log in to customize this list](#))



Human

NCBI36



Mouse

NCBIM37



Zebrafish

New to Ensembl?

Did you know you can:

 [Learn how to use Ensembl](#)

with our video tutorials and walk-throughs

 [Add custom tracks](#)

using our new Control Panel

 [Upload your own data](#)

and save it to your Ensembl account

 [Search for a DNA or protein sequence](#)

using BLAST or BLAT

 [Fetch only the data you want](#)

from our public database, using the Ensembl Perl API

 [Download our databases via FTP](#)

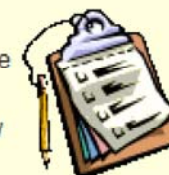
in FASTA, MySQL and other formats

 [Mine Ensembl with BioMart](#)

and export sequences or tables in text, html, or Excel format

User Survey

Almost 6 months have passed since the release of the new website design. If you have a few minutes to spare, we would love to hear what you think of it:



- [Take the survey...](#)

Manage your data

Export data

Bookmark this page

4 Genes match your query ('PDK1') in Human

PDK1

Description

pyruvate dehydrogenase kinase, isozyme 1 [Source:HGNC Symbol;Acc:8809] [Type: protein coding Ensembl/Havana merge]

Gene ID

ENSG00000152256

Location

[2:173420101-173489823:1](#)

Variations

[Variation Table](#)

Source

e67

PDPK1

Description

3-phosphoinositide dependent protein kinase-1 [Source:HGNC Symbol;Acc:8816] [Type: protein coding Ensembl/Havana merge]

Gene ID

[ENSG00000140992](#)

Location

[16:2587965-2653189:1](#)

Variations

[Variation Table](#)

Source

e67

CENPI

Description

centromere protein I [Source:HGNC Symbol;Acc:3968] [Type: protein coding Ensembl/Havana merge]

Gene ID

[ENSG00000102384](#)

Location

[X:100353178-100418670:1](#)

Variations

[Variation Table](#)

Source

e67

C11orf51

Description

chromosome 11 open reading frame 51 [Source:HGNC Symbol;Acc:24531] [Type: protein coding Ensembl/Havana merge]

國際網路 | 受保護模式: 關閉

100%

Step 2 :



Step 1 : copy and past ENSG ID

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Human (GRCh37) Location: 2:173,420,101-173,489,823 Gene: PDK1

Gene: PDK1 **ENSG00000152256**

Description pyruvate dehydrogenase kinase, isozyme 1 [Source:HGNC Symbol;Acc:8809]
Location [Chromosome 2: 173,420,101-173,489,823](#) forward strand.
Transcripts This gene has 11 transcripts

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
PDK1-001	ENST00000282077	1989	ENSP00000282077	436	Protein coding	CCDS2250
PDK1-002	ENST00000410055	2987	ENSP00000386985	436	Protein coding	CCDS2250
PDK1-003	ENST00000392571	1515	ENSP00000376352	456	Protein coding	-
PDK1-004	ENST00000443353	842	ENSP00000399558	154	Protein coding	-
PDK1-007	ENST00000439519	558	ENSP00000388366	154	Protein coding	-
PDK1-008	ENST00000416991	699	ENSP00000399160	229	Protein coding	-
PDK1-201	ENST00000543905	1748	ENSP00000438567	360	Protein coding	-
PDK1-202	ENST00000544863	1338	ENSP00000437502	281	Protein coding	-
PDK1-005	ENST00000436490	571	ENSP00000406608	62	Nonsense mediated decay	-
PDK1-006	ENST00000431718	728	ENSP00000413629	164	Nonsense mediated decay	-
PDK1-009	ENST00000466437	599	No protein product	-	Retained intron	-

Transcript and Gene level displays

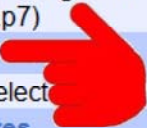
In Ensembl we provide displays at two levels:

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors Login · Register

[New](#) [Count](#) [Results](#) [★ URL](#) [XML](#) [Perl](#) [Help](#)

Dataset

Homo sapiens genes (GRCh37.p7)

Filters 

[None selected]

Attributes

Ensembl Gene ID
Ensembl Transcript ID

Dataset

[None Selected]

Step 1

Ensembl Genes 67

Step 2

Homo sapiens genes (GRCh37.p7)

Step 3

完成 國際網路 | 受保護模式: 關閉 100%


[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

[New](#) | [Count](#) | [Results](#)

[URL](#) | [XML](#) | [Perl](#) | [Help](#)

[Login](#) · [Register](#)

Dataset

Homo sapiens genes (GRCh37.p7)

Filters

Ensembl Gene ID(s) [e.g. ENSG00000139618]: [ID-list specified]

Attributes

Ensembl Gene ID

Ensembl transcript ID

Data

[None Selected]

Please restrict your query using criteria below

REGION:

GENE:

Limit to genes ...

with Illumina HumanWG 6 v1 probe ID(s)

Only

Excluded

ID list limit

Ensembl Gene ID(s) [e.g. ENSG00000139618]

ENSG00000152256

Transcript count >=

Gene type

3prime_overlapping_ncrna

antisense

IG_C_gene

IG_C_pseudogene

IG_D_gene

IG_J_gene

15


[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

[New](#) | [Count](#) | [Results](#)

[URL](#) | [XML](#) | [Perl](#) | [Help](#)

[Login](#) · [Register](#)

Dataset

Homo sapiens genes (GRCh37.p7)

Filters

Ensembl Gene ID(s) [e.g. ENSG00000139618]: [ID-list specified]

Attributes

Ensembl Gene ID
 Ensembl Transcript ID
 Flank-coding region (Gene)
 Upstream flank [2000]

Dataset

[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

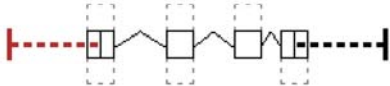
☐ Features
 ☐ Homologs

☐ Structures
 ☐ Variation

☐ Transcript Event
 ☒ Sequences

SEQUENCES:

Sequences (max 1)



☐ Unspliced (Transcript)
 ☐ 5' UTR

☐ Unspliced (Gene)
 ☐ 3' UTR

☐ Flank (Transcript)
 ☐ Exon sequences

☐ Flank (Gene)
 ☐ cDNA sequences

☐ Flank-coding region (Transcript)
 ☐ Coding sequence

☒ Flank-coding region (Gene)
 ☐ Protein

Upstream flank

☒ Upstream flank 2000

Downstream flank

Step 1

Step 2

Step 3


[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

[New](#) | [Count](#) | [Results](#)

[URL](#) | [XML](#) | [Perl](#) | [Help](#)

[Login](#) · [Register](#)

Dataset 1 / 57945 Genes

Homo sapiens genes (GRCh37 p7)

Filters

Ensembl Gene ID(s) [e.g. ENSG00000139618]: [ID-list specified]

Attributes

Ensembl Gene ID

Ensembl Transcript ID

F flank-coding region (Gene)

Upstream flank [2000]

Dataset

[None Selected]

Structures

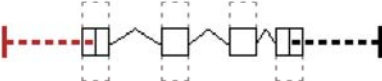
Variation

Transcript Event

Sequences

SEQUENCES:

Sequences (max 1)



Unspliced (Transcript)

Unspliced (Gene)

F flank (Transcript)

F flank (Gene)

F flank-coding region (Transcript)

F flank-coding region (Gene)

5' UTR

3' UTR

Exon sequences

cDNA sequences

Coding sequence

Protein

Upstream flank

☒ Upstream flank 2000

Downstream flank

☐ Downstream flank

Header Information

完成

 網際網路 | 保護模式: 關閉

100%

 [T/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#) Login · Register

[New](#) [Count](#) [Results](#) [★ URL](#) [XML](#) [Perl](#) [Help](#)

Dataset 1 / 57945 Genes

Homo sapiens genes (GRCh37.p7)

Filters

Ensembl Gene ID(s) [e.g. ENSG00000139618]: [ID-list specified]

Attributes

Ensembl Gene ID
Ensembl Transcript ID
Flank-coding region (Gene)
Upstream flank [2000]

Dataset

[None Selected]

Export all results to ☐ Unique results only

Email notification to

View rows as ☐ Unique results only

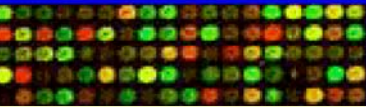
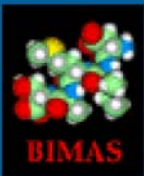
>ENSG00000152256|ENST00000544863;ENST00000543905;ENST00000466437;ENST00000416991;ENST00000439519;ENST0000028
GCAGCAAAATTGCAAAACAGGGATTGAGGGAGCTGAATAATGCTGTGCATGATAAAGTGAC
AAGACTTTGGAGCTCAGGGAATAATCTAGAAGGAAGGAGATTAGAAAGGAGGAGGCAACC
TGAGAGGCTGGGGTAGAGCTGTGGAATGTTTGAAGGACTATTTTAAGAATGTGTAATAT
TCAACTTAAAATTGTTTTCTTTCCAACTTAGGTCCCGCAGAAGAAGGGTGGGAAGAA
GAAGGGGGCCATTGAGCCATCAATGAGGTGGTGACACCATCAACATTCCCAAGCACATCC
CATTGAAACGGACTTCAGGAAGTGTGCCCTCGGGCACTCAGCAAAGATATCTGGAAATT
TGCCAAGAAAGAGATGGGCACTCCAGATGCGCATACTGATACCAGGTTTGACAAAGCTGT
CTGGCCAAAAGAATAATGTCCCATACGGTATCCACATGTGGTTGTCCAGAAAAACATAAT
GCGAATGAAGATTCAGTATATGTGAAAGGAAAATATCACGGGCCACAAAATCACTGAGC
TAAAGGGAAAATTCAAGCTGGGAACGTCTCAAGGCAAACCTGCCTCCCATTCTATTCAAA
GTCATCCCTCTGCTGAGACAGATGCATATCTGATTGCTTCCTTGGGAAAGGCTTATCAGA
AACTCCAAAGACTGCAACCATTTGTCTCTCACCTACTTATAACCTGGAAGCCTTCCTCCC
TGCTTCAAGTTGTCACCTTTCCGGCAGGAACCAATGTACTTCTTGTCATATAGATTGAG
GTCTCCTGTCTCCCTGTATAAAACCAAGCTGTTCAACCTTGGGCACGTGTTCTCAGGAC
TTCCTGAGGCTGTGTACGGGCACGCTCCTCAACCTTGTCAAGATAAACTTTATACATT
AACTGAGACTTGTCTCAAACCTTCGGAGTTCAAAAACATGTACCCGTTACACTTTCTA
AAACCAACAGTCAGTGTGGGAGAACTGCTGATACTCGAATACATCAATAGTTATAAAAC

完成

網路網路 | 受保護模式: 關閉 100%

Signal Scan

WEB ▼ BBS ▼ <http://thr.cit.nih.gov/molbio/signal/> Google



BioInformatics and Molecular Analysis Section
Computational Bioscience and Engineering Lab, Division of Computational Bioscience
Center for Information Technology, National Institutes of Health

WWW Signal Scan

Function: Find and list homologies of published signal sequences with the input DNA sequence.
The [analysis](#) is done using the SIGSCAN Version 4.05 suite of programs developed by [Dr. Dan Prestridge](#), Information on SIGSCAN is maintained at the [Advanced Biosciences Computing Center](#), University of Minnesota.. This WWW Signal Scan is based on WEB version implemented by: [Meena Sakharkar](#) at [Biocomputing Research and User Support Technology Group](#), [National University of Singapore](#).

TRANSFAC ▼ [Signal Database](#) source (Use this [form](#) for IMD Signal Search)

Select signal classes to scan (at least one):

☐ Mammal	☐ Bird	☐ Amphibian	☐ Prokaryote	☐ Virus (<i>TRANSFAC only</i>)
☐ Insect	☐ Plant	☐ Other Eukaryotes	☐ Yeast (<i>TFD only</i>)	

<http://www-bimas.cit.nih.gov/molbio/signal/>

TRANSFAC

[Signal Database](#) source (Use this [form](#) for IMD Signal Search)

Select signal classes to scan (at least one):

- ☒ Mammal ☐ Bird ☐ Amphibian ☐ Prokaryote ☐ Virus (*TRANSFAC only*)
☐ Insect ☐ Plant ☐ Other Eukaryotes ☐ Yeast (*TFD only*)

Display the Results: ☒ [grouped by signal](#) ☐ [by sequence order](#) ☐ [mapped to sequence](#)Please enter or paste a Nucleic Acid sequence to analyze (most [formats](#) accepted):

```
TGGGACAATAGTTACCCATTGTATTCAAAAGTAATGAAAACTAATAAAGTGCAG
ATAGGAAAAAATTAATTTTAGCCAG
AAGATACATGAAACATGTCAGCAGATACACGTGTGCCAGTCTGTAAAGCAAGACC
ACTTTTTGGTTCAGCCCTATTTGTG
GTTTGTCATTTCACTCTATTGGTTGGTTCCAGAATTTAGGAAGTTACTAGT
AACTTTGCAAGGCTAGTTCAGGCAG
```

☒ Echo input sequence (generally [recommended](#))

copy & paste

submit

reset

Credits: WWW implementation by [Robin Hart](#) & [Rao Parasa](#)

If you use this program in published research, please cite:

WEB		BBS		http://thr.cit.nih.gov/cgi-bin/molbio/signal		Goo	
GT-1	site	1065	(-)	nnCCACCnnnn	R03492	BioBase Login required	
H1TF2	fact	1219	(+)	CCAAT	R00660	BioBase Login required	
H1TF2	fact	178	(-)	CCAAT	R00660	BioBase Login required	
H4TF-1	fact	298	(+)	GATTTC	R00680	BioBase Login required	
H4TF-2	fact	1269	(+)	GGTCC	R00681	BioBase Login required	
H4TF-2	fact	1514	(-)	GGTCC	R00681	BioBase Login required	
H4TF-2	fact	389	(+)	GGTCC	R00681	BioBase Login required	
H4TF-2	fact	751	(-)	GGTCC	R00681	BioBase Login required	
HNF-1	fact	673	(-)	ATTAAC	R00436	BioBase Login required	
HNF-3	fact	580	(+)	GTTTGTITT	R04324	BioBase Login required	
II	site	767	(-)	CTGGTCAG	R03193	BioBase Login required	
input.seq: 2000 base pairs							
IRF-2	fact	453	(-)	AAGTGA	R00917	BioBase Login required	
LF-A1	fact	1770	(-)	GGGCA	R01171	BioBase Login required	
LF-A1	fact	334	(-)	GGGCA	R01171	BioBase Login required	
LF-A1	fact	565	(-)	GGGCA	R01171	BioBase Login required	
LF-A1	fact	713	(+)	GGGCA	R01171	BioBase Login required	
MAZ	fact	1818	(-)	GGGAGGG	R02306	BioBase Login required	
myogenin	fact	710	(-)	TGCCCTTG	R00018	BioBase Login required	
myogenin	fact	789	(-)	TGCCTGG	R00018	BioBase Login required	
NF-1	fact	1219	(+)	CCAAT	R01543	BioBase Login required	
NF-1	fact	178	(-)	CCAAT	R01543	BioBase Login required	
NF-1	fact	1052	(+)	GCCA	R01681	BioBase Login required	
NF-1	fact	1066	(-)	GCCA	R01681	BioBase Login required	
NF-1	fact	115	(+)	GCCA	R01681	BioBase Login required	
NF-1	fact	1158	(-)	TCCA	R01681	BioBase Login required	

TFSEARCH

Japanese

[Notice to TFSEARCH users](#)
[TFFACTOR \(sorted by mol.weight, partial\)](#)

Since 01/Apr/98 02946009

[Help](#)

Enter a label for the sequence:

完成

網際網路 | 受保護模式: 關閉

100%

TESS String search

TESS - String Search Page

[Home](#) [AnGEL CRM Searches](#) [Site Searches](#) [Query Databases](#) [Other Stuff](#)

[GuideRecall SearchCombinedStringsFiltered Strings](#)

Welcome 58.114.91.6

Need help? Check our [FAQ page](#) then please send questions and comments to TessMaster@cbil.upenn.edu.

Please see [this](#) announcement for important information about site searches performed between early July and December 17, 2007.

What potential transcription factor binding sites are there in my sequence?

Use this form to enter your DNA sequences, select which database of TFBS you want to match against your sequence and/or enter your own sites or PWMs.

Input

Enter the minimal information needed to submit a job to TESS.

Title: (text)

You may submit multiple sequences in this window. Each sequence can have a maximum length of 2000[bp]. The sequences must a total length less than 100000 [bp].

DNA Sequence(s):

(text)

Length of time to store results of the job:

Your email address: (string)

End of Minimal Parameters

PROMO

Google less string search Search More >> shih-Jay Lin

Y! tess string search 搜尋

Gmail - What's new for 'C... Ensembl genome brows... ALGGEN - PROMO

RESEARCH

- PROMO**
A virtual laboratory for the study of transcription factor binding sites in DNA sequences.
- Factor's species:
All species
- Site's species:
All species
- [SelectSpecies](#)
- [SelectFactors](#)
- [ViewMatrices](#)
- [SearchSites](#)
- [MultiSearchSites](#)
- [Help](#)

[Transcription factors](#) See information on the transcription factors from TRANSFAC.

PROMO HOME PAGE

PROMO is a virtual laboratory for the identification of putative transcription factor binding sites (TFBS) in DNA sequences from a species or groups of species of interest. TFBS defined in the TRANSFAC database are used to construct specific binding site weight matrices for TFBS prediction. The user can inspect the result of the search through a graphical interface and downloadable text files.

PROMO is using version 8.3 of TRANSFAC.

If you use the results in your research please cite:

Xavier Messeguer, Ruth Escudero, Domènec Farré, Oscar Nuñez, Javier Martínez, M.Mar Albà. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics*, 18, 2, 333-334, 2002.

Domènec Farré, Romà Roset, Mario Huerta, José E. Adsuara, Llorenç Roselló, M.Mar Albà, Xavier Messeguer. Identification of patterns in biological sequences at t and MALGEN. *Nucleic Acids Res*, 31, 13, 3651-3653, 2003.

Search for putative sites in your sequence:

step 1.

[SelectSpecies](#): select the species or group of species of interest.

Current factor's species or group: **All species**
Current site's species or group: **All species**

Also you can [SelectFactors](#) to restrict the prediction to a transcription factor set.

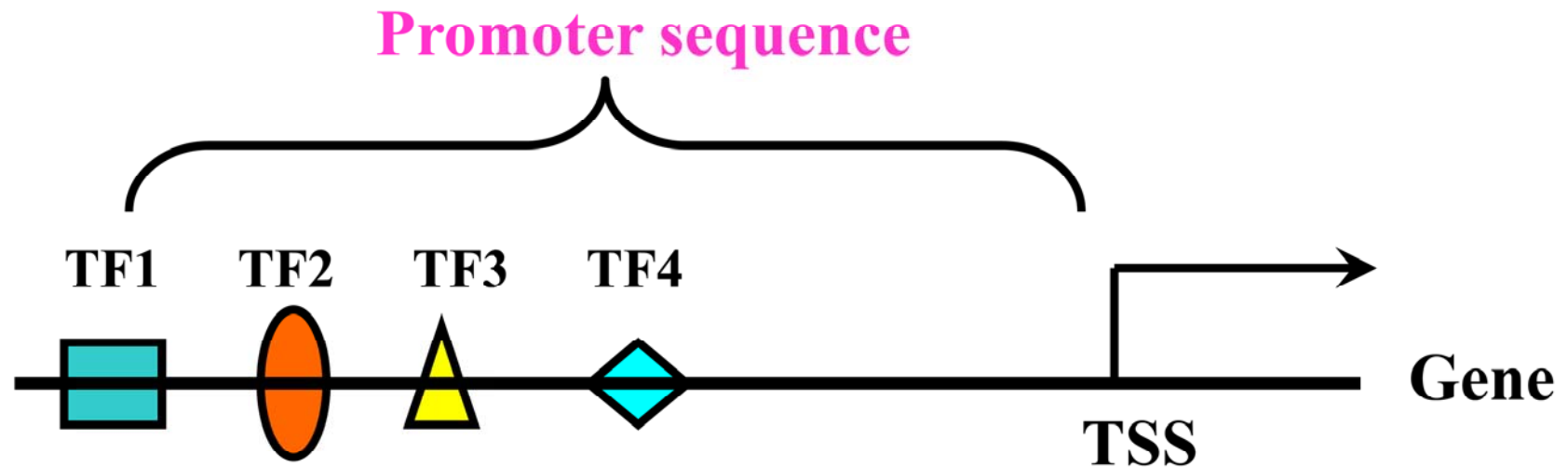
step 2

網際網路 | 受保護模式: 關閉 100%

練習

選擇一個你有興趣的基因分析其promoter 是否可被 CREB 調控？並列出其core 及 flanking sequence。

Predict transcription factor binding element by bioinformatics tools



- **Signal Scan**

<http://www-bimas.cit.nih.gov/molbio/signal/>

- **TFSEARCH**

<http://www.cbrc.jp/research/db/TFSEARCH.html>

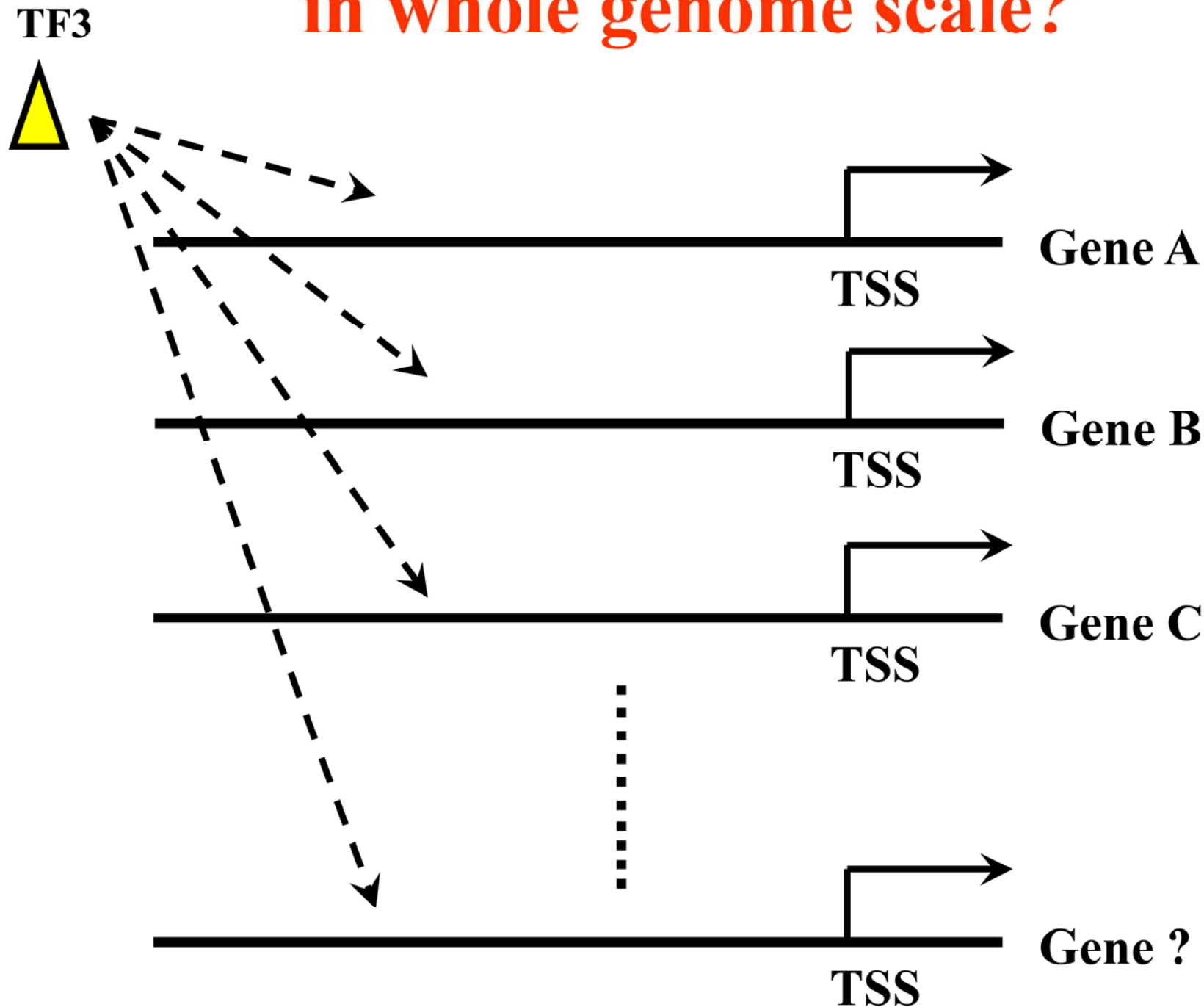
- **TESS**

<http://www.cbil.upenn.edu/cgi-bin/tess/tess33?RQ=SEA-FR-QueryS>

- **PROMO**

http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3

How to predict potential target gene of TF in whole genome scale?



Part III:

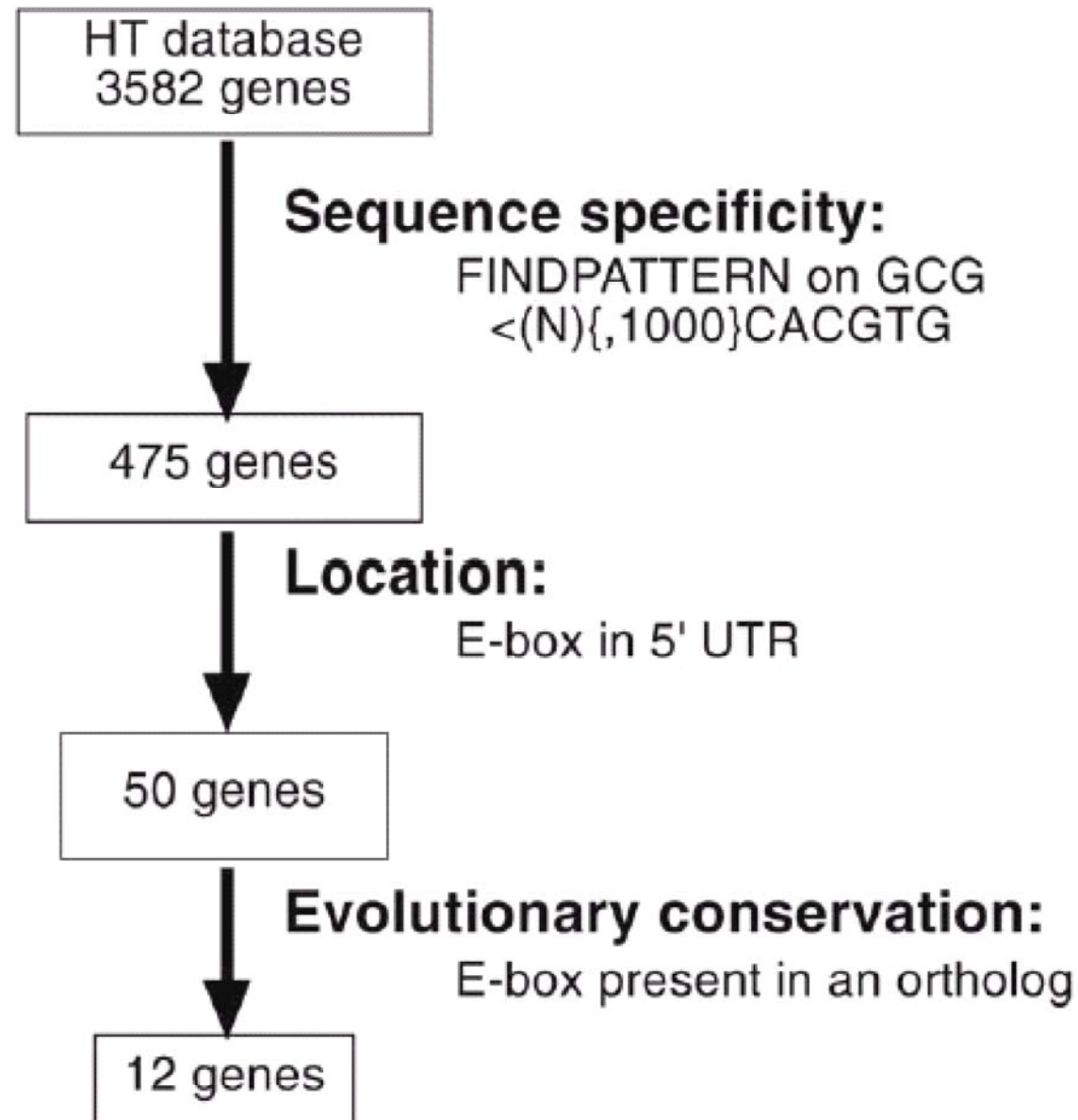
Whole Genome Search of Transcription Factor Regulated Candidate Genes by Profile Hidden Markov Model



Identify transcription factor binding sites

- **Consensus search**
- **Position Specific Scoring Matrix (PSSM)**
- **Hidden Markov Model**

Consensus



PSSM (TRANSFAC)

A multiple alignment based motif/domain/element

A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C

	1 _{st}	2 _{nd}	3 _{rd}	
A	.8	0	.8	
G	0	.2	0	
C	0	.8	.2	
T	.2	0	0	

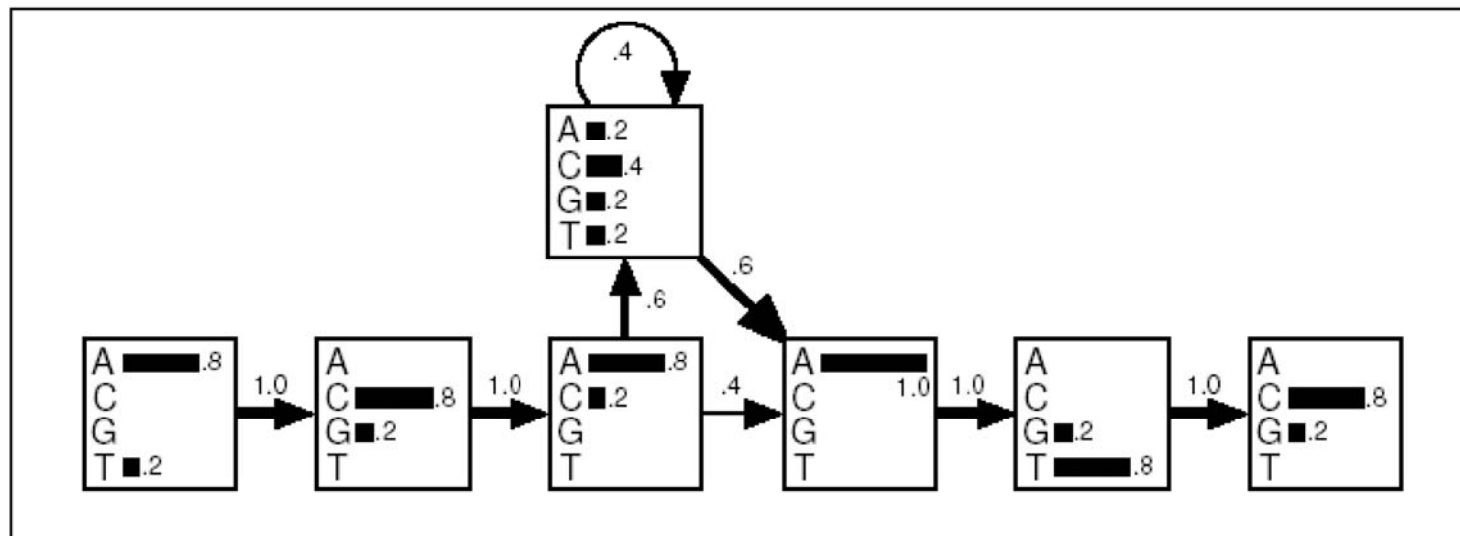
→ Length of gap?

Computational Methods in Molecular Biology, p45-63. Elsevier, 1998.

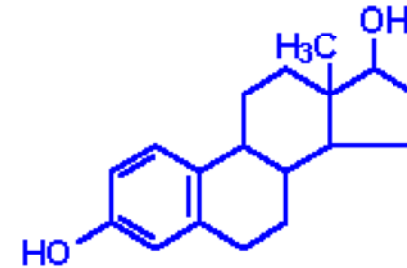
Hidden Markov Model

A multiple alignment based motif/domain/element

A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C

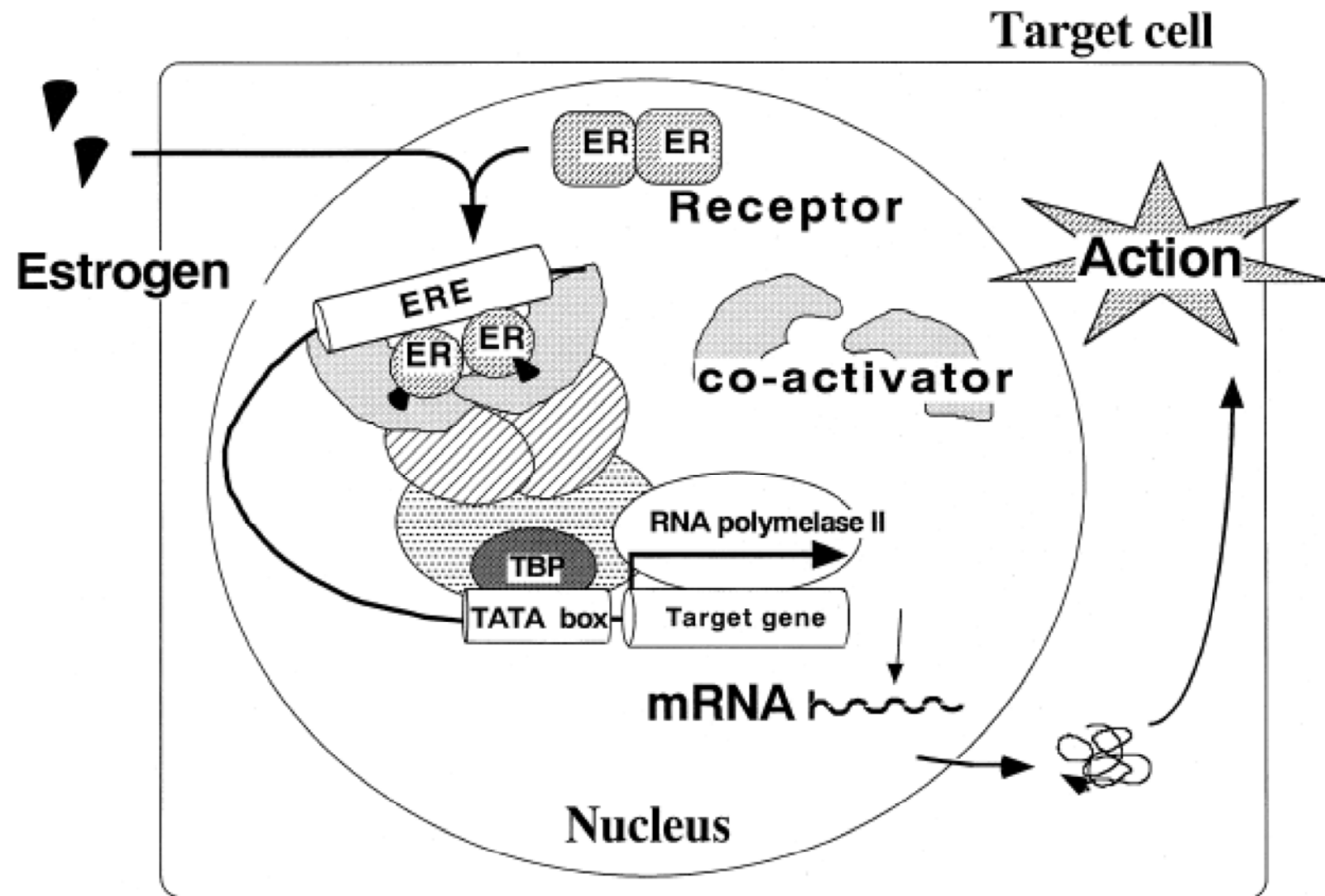


17b-Estradiol



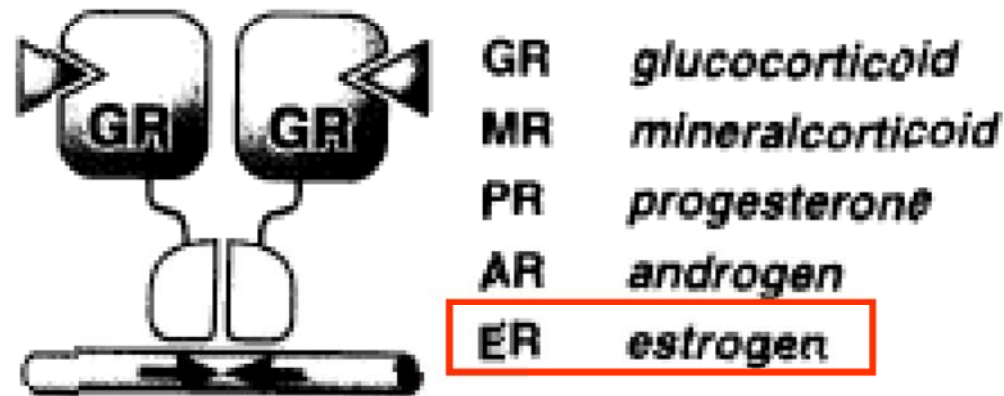
- **an estrogen, principal female sex hormone, produced in the ovary, responsible for secondary female sex characteristics**

Pathway of the Estrogen Receptor (ER)



Estrogen Response Element (ERE)

Steroid Receptors



palindromic

→ ←
GGTCAnnnTGACC
CCAGTnnnACTGG

The Binding Element Searching Tool (THE BEST)

- **Users could build up a customized **hidden Markov model (HMM)** via HMMER program by upload or paste a set of element sequences containing specific transcription factor binding site (in FASTA format) and use this model to search the multi-species promoter region (human, mouse and rat) retrieved from the EnsMart system annotated by ensembl project**

Motivation

- Search for novel estrogen regulated target genes by known ERE based on multiple alignments

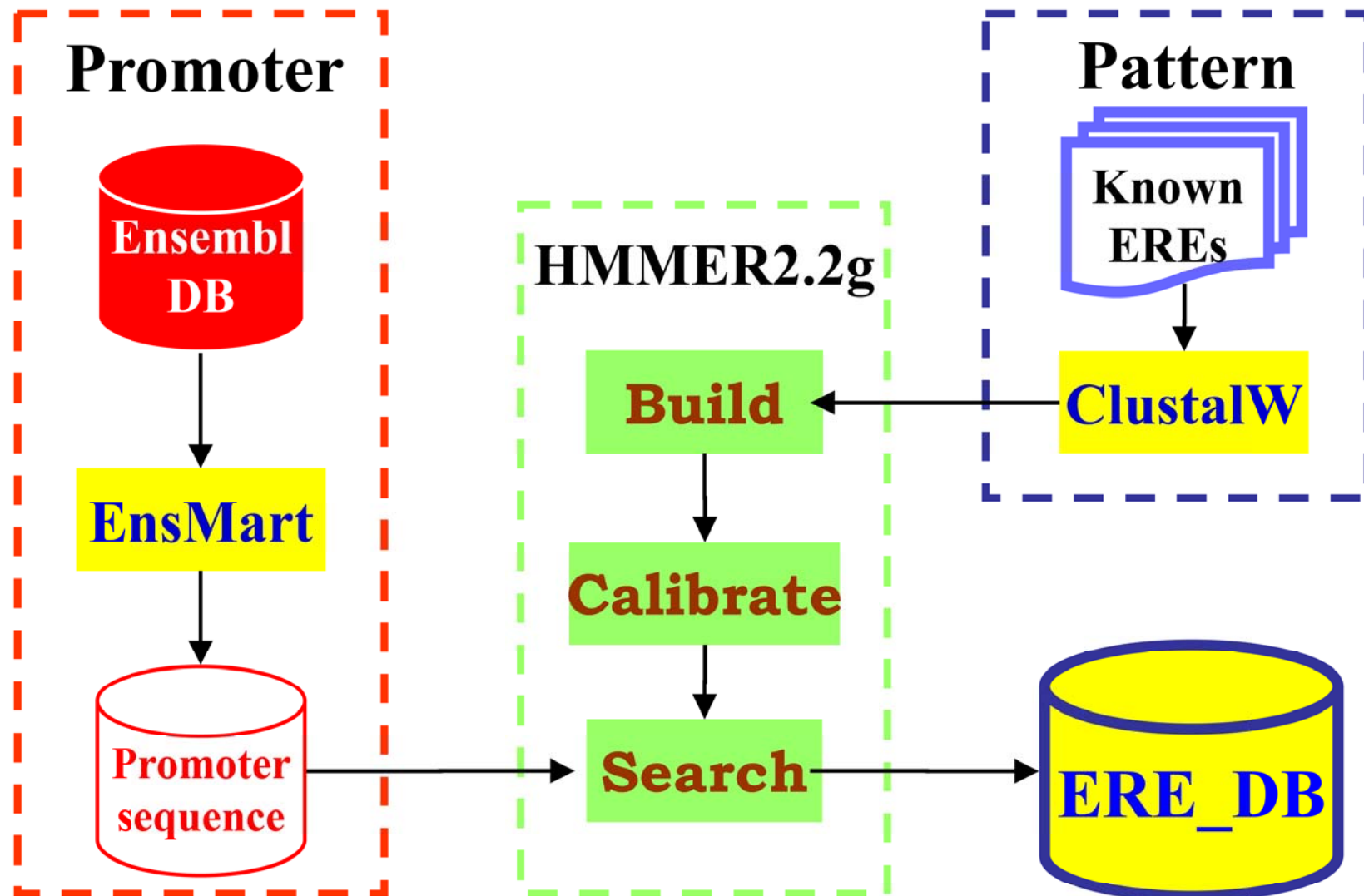


Literature search

(promoter activity 、 gel shift assay 、 ChIP)

System modules

Hidden Markov Model based search tools: HMMER2.2g



◎應用資源

[Vector NTI](#)
[IBM軟體下載](#)
[NCKU BLAST](#)
[NCKU SNP Database](#)
[Repeat Masker Server](#)
[TAG Database](#)
[TheBEST](#)

本中心開授 非同步遠距教學課程 如下：

正規課程：

- ◎ 生物資訊學
- ◎ 生物資訊學導論
- ◎ 分子生物學
- ◎ 高等生理學
- ◎ 電腦在生物相關資源上的應用
- ◎ 分子生物學概論
- ◎ 遺傳學特論

暑期課程：

- ◎ 生物資訊理論與實務應用
- ◎ 基因體學
- ◎ 生物資訊在醫學上之應用
- ◎ 網頁應用及程式設計
- ◎ 整合性生物資訊分析軟體之應用

研習會：

- ◎ 生物資訊分析軟體教育課程(Vector NTI)
- ◎ JEMBOSS基礎教育課程
- ◎ ABI 7900儀器設計原理及軟體操作

NCKU Bioinformatics Center

THE Binding Element Searching Tool

● Introduction

● Start

● My Result

● Logout

STEP 1. Multiple Sequence Alignment	
a.	Enter a set of sequences in FASTA format:
	>vit-ERE CGTCAGGTCACAGTGACCTGATG >fos-ERE ATCCAGGTCACACAGCCCAGGC >jun-ERE-1 CTGAAGCAGAGCATGACCTTGAA
	Or upload a file in FASTA:
b.	I want using my pre-aligned seq. in MSF format:
NEXT	

NCKU Bioinfo

THE Binding

Introduction

Start

My

STEP 1. Multiple Sequence Alignment

Done!

View [result of multiple sequence alignment](#).

STEP 2. Construct and calibrate a HMM

- ☒ global in HMM, local in sequence
- ☐ multiple local alignments (-f)
- ☐ single global alignment (-g)
- ☐ single local alignment (-s)

BACK

NEXT

```
1 25
ENSG000000138061 CTGCCAGGTCGCGGTGCCCTCCTTC
ENSG000000160182 CTGCAAGGTCACGGTGGCCACCCCG
ENSG000000087510 ACGCCTGGTCACCGTGACCCCGATT
ENSG000000115944 TGCAGGGGTCAAGGTGACCCCGGG
ENSG000000085465 TTCTGGGGTCACTGTGACTCTCATA
ENSG000000147654 TTTCCGGGTCAAGGTGACCTCTGGG
ENSG000000117984 AAGCTGGGCCGGGTGACCCGCGG
ENSG000000164362 CATGTTGGTCAGGTGATCTCAAAA
ENSG000000135744 TAAATAGGGCATCGTGACCCGGCCG
ENSG000000118777 GTGTCACGGCAGGTGACCTAGCC
ENSG000000171105 TACACAAGTGAACTGGCCAGAGA
```


NCKU Bioinform

THE Binding E

Introduction

Start

My Res

STEP 1. Multiple Sequence Alignment

Done!

View [emma result](#).

STEP 2. Construct and calibrate a HMM

Done!

View [hmmbuild result](#), [hmmcalibrate result](#), [hmm](#)

STEP 3. Search the specific HMM in a dataset

sets T bit threshold (globT) to

a.

- ☐ Human
☐ Mouse
☐ Rat

Because this step needs some time to complete, please remember your account: **q26pspivpv** for

http://thebest.binfo.ncku.edu.tw/thebest/usr/process/view.php?display=swos_result

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
ENSG00000115944	1/1	6	25 .]	1	19 []	9.2	0.0018
ENSG00000147654	1/1	6	25 .]	1	19 []	7.5	0.0045
ENSG00000087510	1/1	6	25 .]	1	19 []	7.1	0.0057
ENSG00000160182	1/1	6	25 .]	1	19 []	6.5	0.0079
ENSG00000117984	1/1	6	25 .]	1	19 []	5.6	0.013
ENSG00000135744	1/1	6	25 .]	1	19 []	4.5	0.023
ENSG00000085465	1/1	6	25 .]	1	19 []	3.8	0.033
ENSG00000164362	1/1	6	25 .]	1	19 []	3.7	0.037
ENSG00000118777	1/1	6	25 .]	1	19 []	3.6	0.038
ENSG00000138061	1/1	6	25 .]	1	19 []	2.4	0.074
ENSG00000171105	1/1	6	25 .]	1	19 []	-0.3	0.32

Alignments of top-scoring domains:

ENSG00000115944: domain 1 of 1, from 6 to 25: score 9.2, E = 0.0018

->aggtcaggggtgaccccgga.g<-

+gggtca ggtgacccc++g

ENSG0000001 6 GGGTCAAGGTGACCCCGgG 25

ENSG00000147654: domain 1 of 1, from 6 to 25: score 7.5, E = 0.0045

->aggtcaggggtgaccccgga.g<-

+gggtcaggggtgacc+c ++g

ENSG0000001 6 GGGTCAGGGTGACCTCTGgG 25

NCKU Bioinformatics Center

THE Binding Element Searching Tool

● Introduction ● Start ● My Result ● Logout

	Enter your account:	
	q26psjvpv	
	NEXT	

Practice account: [hedpqst3oh_](#)

NCKU Bioinformatics Center

THE Binding Element Searching Tool

● Introduction ● Start ● My Result ● Logout

	Select:		
	☒ Human		
	☐ Mouse		
		BACK	NEXT



Sections of Result Access

Human menu:	
a.	View all GO Clustering
b.	Field searching
you can use "" or - Gene Description: GO Description: Search	
c.	ENSG searching
ENSG list: Search	
d.	AFFY ID searching
AFFY ID list: Search	
BACK	

See raw data or cluster result by gene ontology term (level :3)

Retrieving user-known genes

1. By gene name/descriptions ex: FGF, growth factor
2. By Gene Ontology term/Key words ex: growth, metabolism

Retrieving subset of result by known ENSG ID (list)

ex: ENSG00000129991

Retrieving subset of result by known AFFYMETRIX microarray ID (list)

ex: 205742 at

Gene Ontology clustering

Gene Ontology Clustering :

Please Check the type :

☐ Molecular Function ☒ Biological Process ☐ Cellular Component

Submit

By GO, user annotates TF-regulated targets and has an insight in biological function of predicted transcriptome.

fertilization

Click for the details of genes in this category

0.33% [14 / 4262] [genes in category / total annotated genes]

[ENSG00000104450](#) [ENSG00000146839](#) [ENSG00000116996](#) [ENSG00000187191](#) [ENSG00000111254](#)
[ENSG00000118245](#) [ENSG00000158806](#) [ENSG00000113552](#) [ENSG00000104755](#) [ENSG00000103310](#)
[ENSG00000004809](#) [ENSG00000140545](#) [ENSG00000149506](#) [ENSG00000164078](#)

reproductive organismal physiological process

0.33% [14 / 4262] [genes in category / total annotated genes]

[ENSG00000102755](#) [ENSG00000112782](#) [ENSG00000180914](#) [ENSG00000182752](#) [ENSG00000084636](#)
[ENSG00000104870](#) [ENSG00000162244](#) [ENSG00000145708](#) [ENSG00000083782](#) [ENSG00000101405](#)
[ENSG00000087460](#) [ENSG00000087494](#) [ENSG00000007866](#) [ENSG00000189231](#)

THE **B**inding **E**lement **S**earching **T**ool

● Introduction

● Start

● My Result

● Logout

Human menu:

a. [View all](#)

b. Field searching

you can use "" or -

Gene Description:

StAR

GO Description:

Search

c. ENSG searching

ENSG list:

Search

d. AFFY ID searching

AFFY ID list:

THE Binding Element Searching Tool

Introduction

Start

My Result

Logout

BACK

1.

ENSG ID

[ENSG00000159433](#)

Gene

[STARD9](#)

Gene Description

StAR-related lipid transfer protein 9 (StARD9) (START domain- containing protein 9) (Fragment).

[Source:Uniprot/SWISSPROT;Acc:Q9P2P6]

ENST	Score	Hit Seq.	Position	Orientation
ENST00000290607	-0.8	ACCTCAGAATCACCTTGCT	-1297~-1279	forward
ENST00000290607	-4.8	AGGAGAAGCTGACCCTCTT	-1178~-1160	reverse

AFFY ID

[227108_at](#)

[60454_at](#)

[Hs.173871.0.S1_3p_at](#)

GO Characterization

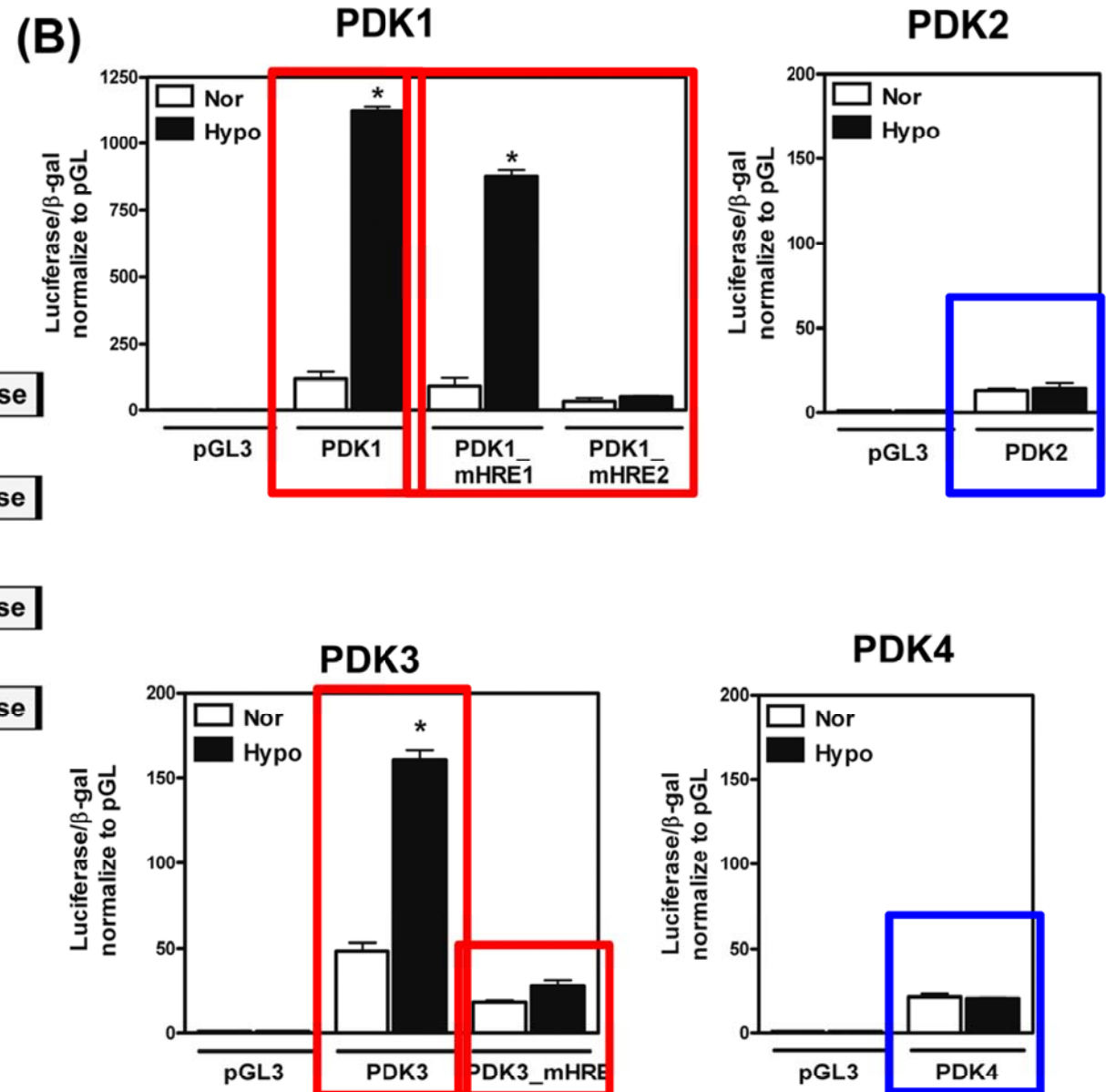
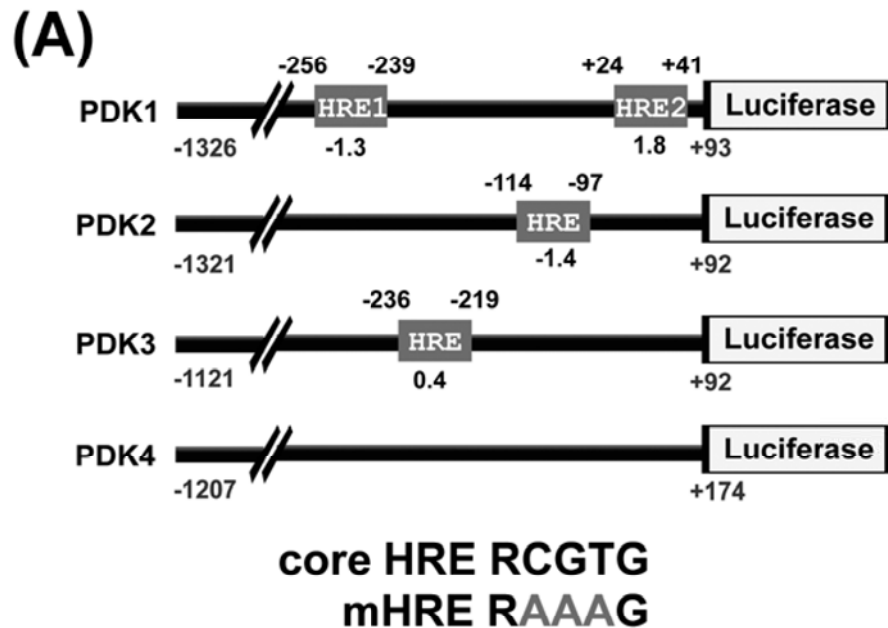
InterPro Characterization

[IPR002913](#)

Lipid-binding START

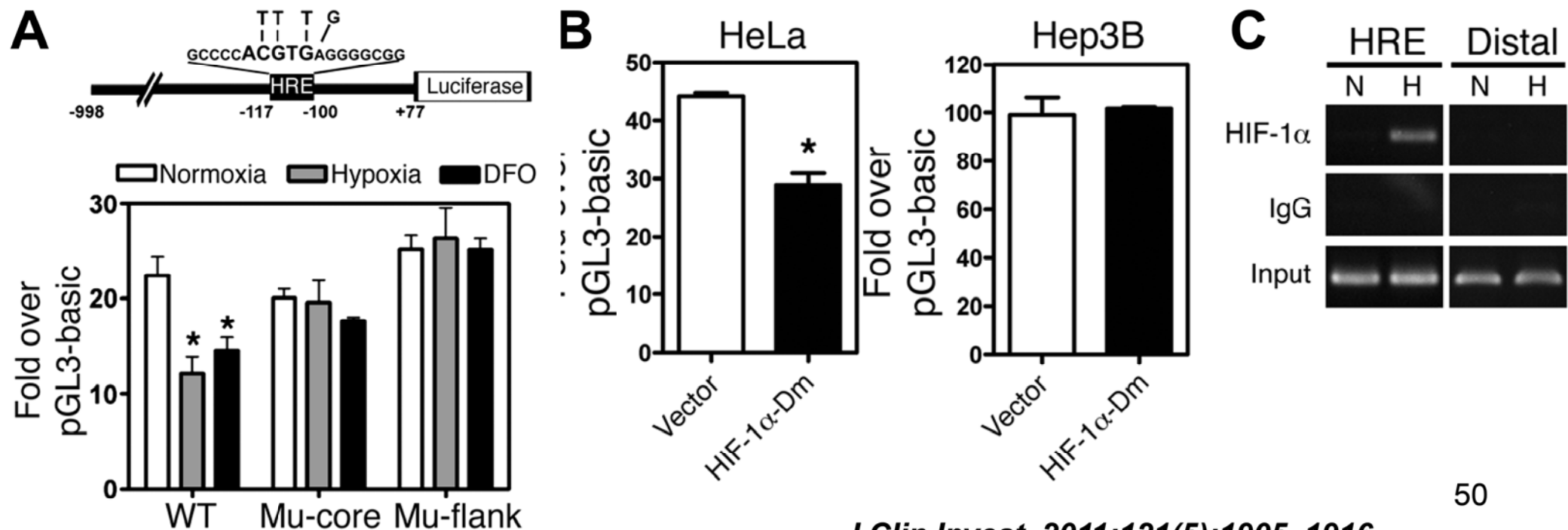
2.

Hypoxia induces promoter activities of PDK1 and PDK3



Hypoxia-repressed DUSP2 expression is mediated by predicted HRE

ENSG ID				
ENSG00000158050				
Gene				
DUSP2				
Gene Description				
Dual specificity protein phosphatase 2 (EC 3.1.3.48) (EC 3.1.3.16) (Dual specificity protein phosphatase PAC-1). [Source:Uniprot/SWISSPROT;Acc:Q05923]				
ENST	Score	Hit Seq.	Position	Orientation
ENST00000288943	0.9	GCCCCACGTGAGGGGCGG	-117~-100	forward



9.

ENSG ID				
ENSG00000147465				
Gene				
STAR				
Gene Description				
Steroidogenic acute regulatory protein, mitochondrial precursor (StAR) (StARD1). [Source:Uniprot/SWISSPROT;Acc:P49675]				
ENST	Score	Hit Seq.	Position	Orientation
ENST00000276449	-3	AGACCAGCATGGCCAACAT	-538~-520	reverse
AFFY ID				
204548_at				
34343_at				
g4507250_3p_at				
GO Characterization				
Component	GO:0005739	mitochondrion		
Function	GO:0008289	lipid binding		
	GO:0015485	cholesterol binding		
	GO:0017127	cholesterol transporter activity		
Process	GO:0006700	C21-steroid hormone biosynthesis		
	GO:0006869	lipid transport		
InterPro Characterization				
IPR002913	Lipid-binding START			
IPR000799	Steroidogenic acute regulatory protein			

練習

- 利用PUBMED search文獻，找出約10個CRE (cAMP-response element binding site) 的sequences (需包含5'與3' flanking sequences，約20bp左右)，以FASTA format 輸入THE BEST進行分析，並列出10個candidate genes

Example

□ 1: [Int J Biochem Cell Biol.](#) 2006;38(10):1808-20. Epub 2006 May 17.

Cyclic AMP stimulates MEG3 gene expression in cells through a cAMP-response element (CRE) in the MEG3 proximal promoter region.

[Zhao J](#), [Zhang X](#), [Zhou Y](#), [Ansell PJ](#), [Klibanski A](#).

Neuroendocrine Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA.

MEG3 is a human maternally expressed gene that potentially acts as a non-coding RNA. Our laboratory found that a cDNA isoform of MEG3, MEG3a, inhibits cell proliferation. MEG3 is highly expressed in the normal human pituitary but not expressed in clinically non-functioning pituitary tumors, suggesting that this imprinted gene may be involved in pituitary tumorigenesis. Previously we demonstrated that hypermethylation of the MEG3 promoter region is associated with the loss of MEG3 expression in pituitary tumors, potentially by blocking the binding of transcription factors to their cis-elements. To further investigate the cis- and trans-factors that are important for the regulation of MEG3, we have characterized the human MEG3 promoter. A single transcription initiation site was identified by 5' RACE. Up to 5kb of the 5'-flanking region of MEG3 gene was cloned into a reporter plasmid. Deletion and mutation analysis suggest that a cAMP response element (CRE), located between -69 and -49 of the MEG3 proximal promoter region, is critical for promoter activity. Consistent with this finding, Northern blot analysis demonstrate that elevated intracellular cAMP levels stimulate MEG3 expression in human fibroblasts in culture. Furthermore, gel shifting, ChIP analysis, and co-transfection experiments show that CREB directly binds to the CRE site and stimulates MEG3 promoter activity. Therefore, MEG3 is a downstream target gene of cAMP. Together with the anti-proliferative function of cAMP, our data suggest that MEG3 may interact with the cAMP-dependent signaling pathway to be involved in the control of cell proliferation and other cAMP-related physiological functions.