

以生物資訊工具確認轉錄因子 及受其調控之目標基因

蕭 貴 陽

國立成功大學
生理學研究所

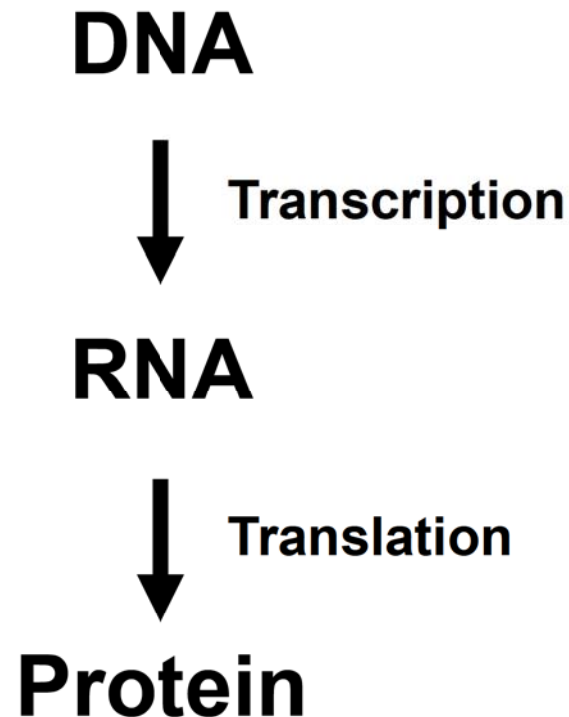
2013-09-06

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**

The Central Dogma of Molecular Biology



Transcriptional Regulation

- 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
 - Gene specific transcription factors

General Transcription Factors

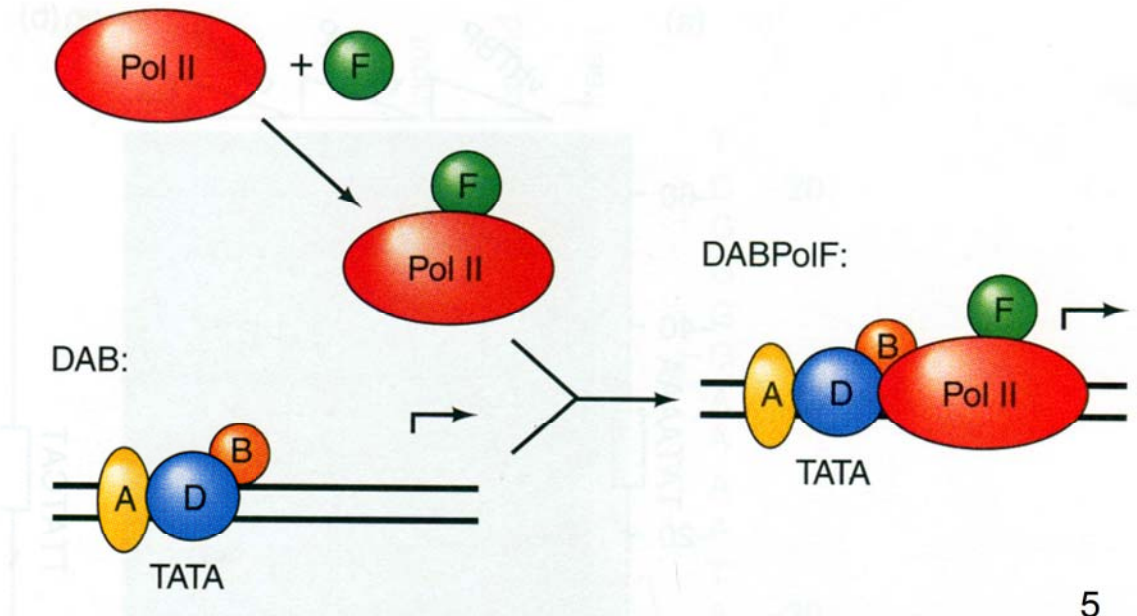
Definition:

bind with RNA polymerase to form a pre-initiation complex only to a weak extent

Class II Factors

TFIID, TFIIA, TFIIB, TFIIF, TFIIE & TFIIH

TBP (TATA box binding protein)

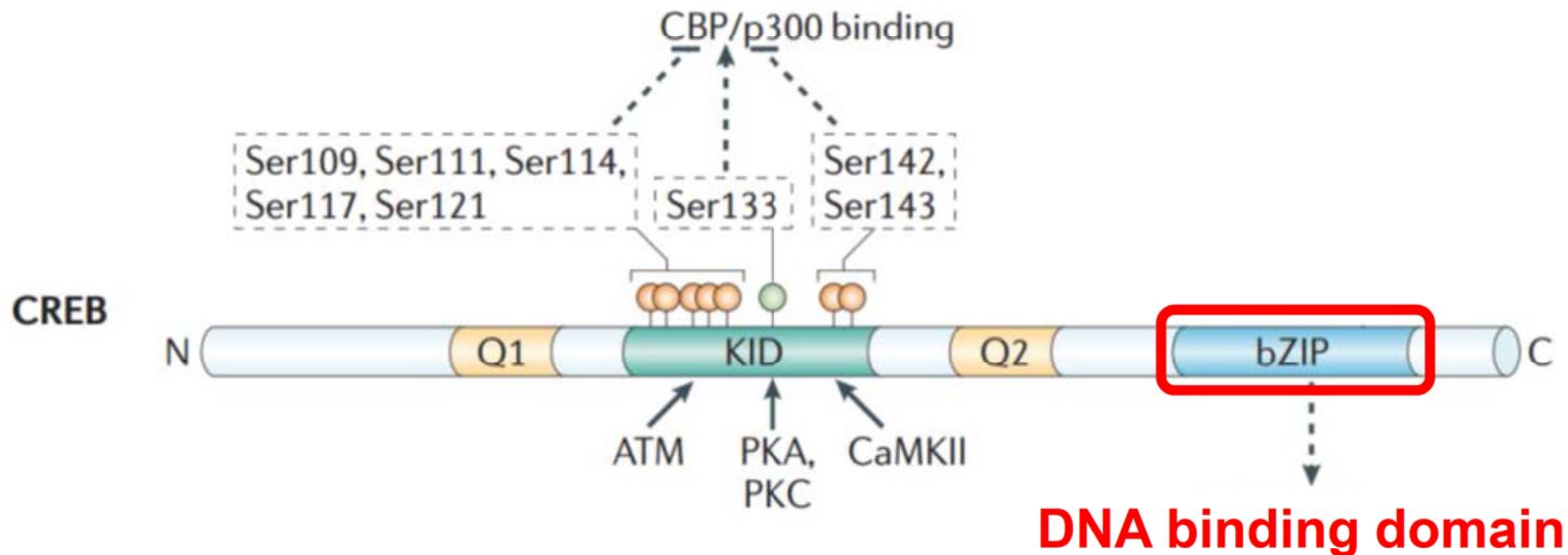


Gene specific transcription factors

- Multiple regulatory sites existed in a gene's promoter – bound by different regulatory proteins
 - N-to-1
- A single regulatory protein recognizes numerous sites in the genome
 - 1-to-N
- Cells respond to environmental stimuli by synthesizing different transcription factors

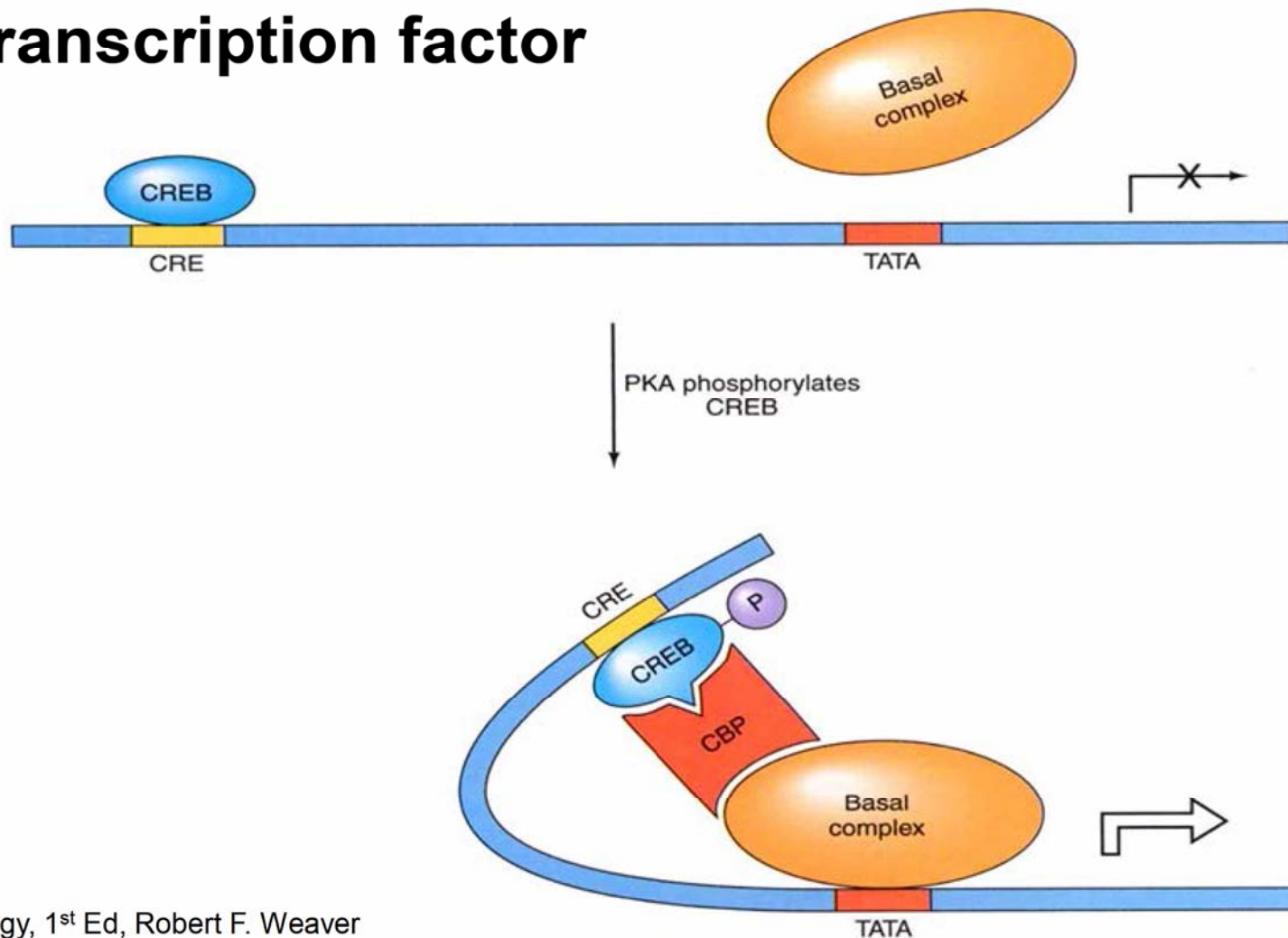
Control of Gene Expression

- **Activation of Transcription**
 - **Transcription factor**



Control of Gene Expression

- **Activation of Transcription**
 - Transcription factor



Transcription Factors

Binding Element

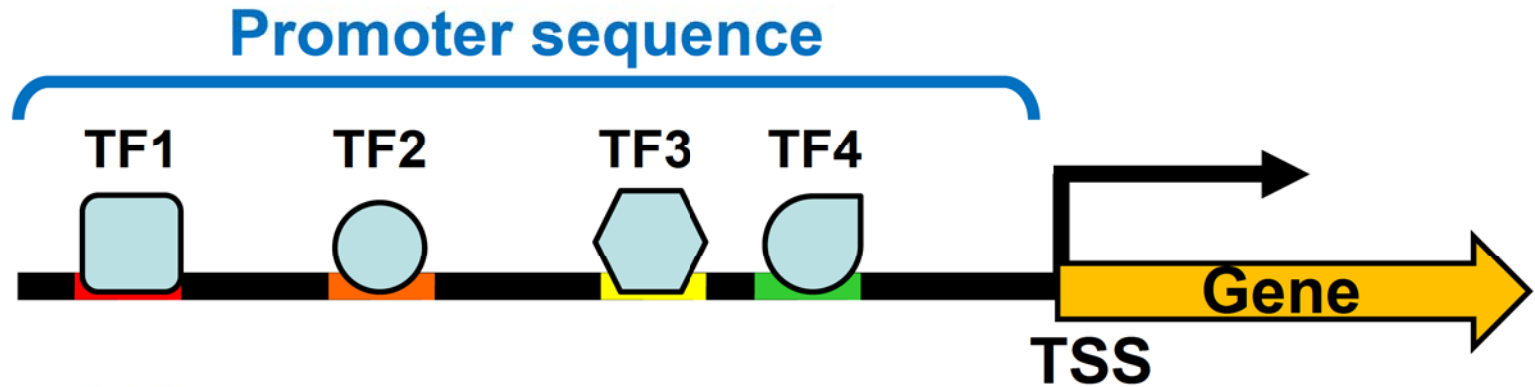
- Example: GRE
 - Glucocorticoid Response Element

5' - AGAACANNNTGTTCT - 3'

3' - TCTTGTNNACAAGA - 5'

- A palindrome
- Two-fold nature is important
- A pair of GR polypeptides binds to DNA

Predict transcription factor binding element by bioinformatic tools



- **Signal Scan**
 - <http://www-bimas.cit.nih.gov/molbio/signal/>
- **TFSEARCH**
 - <http://www.cbrc.jp/research/db/TFSEARCH.html>
- **PROMO**
 - http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3

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

First thing first !

**How to extract promoter
sequence of a specific gene ?**

<http://www.ensembl.org>

Web-based tools used today

- **ENSEMBL & BioMart**
 - Gene ID & sequence retrieval
- **Signal Scan**
 - Position-Specific Weight Matrix-based (PSWM)
 - <http://www.bimas.cit.nih.gov/molbio/signal/>
- **TFSEARCH**
 - PSWM-based
 - <http://www.cbrc.jp/research/db/TFSEARCH.html>
- **PROMO**
 - PSWM-based
 - http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
- **The BEST**
 - HMM-based

Ensembl Genome Browser

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

NCBI - USA
DDBJ - Japan



Ensembl is a joint project between [EMBL - EBI](#) and the [Wellcome Trust Sanger Institute](#) to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.



[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [More ▼](#)

[Login/Register](#)

Search all species...

Search: for

Go

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.

Popular genomes



Human

GRCh37



Mouse

GRCm38

**ENCODE data
in Ensembl**



**Variant Effect
Predictor**



**Gene
expression in
different
tissues**

**Find SNPs
and other
variants for
my gene**

What's New in Release 73 (September 2013)

- [New search engine using Solr](#)
- [New species: Duck \(Anas platyrhynchos\) and Flycatcher \(Ficedulla albicollis\)](#)
- [Updated patches for the human assembly \(GRCh37.p12\)](#)

[Full details of this release](#)

[More release news on our blog →](#)

Latest blog posts

- 04 Sep 2013: [Ensembl 73 has been released!](#)



Search:

Human

for PDK1

Go

All species

Favourite species

Human

Mouse

Zebrafish

Alpaca

Anole lizard

Armadillo

Bushbaby

C.intestinalis

C.savignyi

Caenorhabditis elegans

Cat

Chicken

Chimpanzee

Chinese softshell turtle

Cod

Coelacanth

0000 or coronary heart disease

ENCODE data
in Ensembl



Variant Effect
Predictor



Gene
expression in
different
tissues



Find SNPs
and other
variants for
my gene

GTRTATACATTC
CRTRAAAGTCTT

Retrieve gene
sequence

GCCGTGACTTCGGGGTGG
GGGCTTGTGGCGCAGC
GCGCCTCTGCTGCGCCT

Compare
genes across
species



What's New in Release 73 (September 2013)

- [New search engine using Solr](#)
- [New species: Duck \(Anas platyrhynchos\) and Flycatcher \(Ficedula albicollis\)](#)
- [Updated patches for the human assembly \(GRCh37.p12\)](#)

[Full details of this release](#)

[More release news on our blog →](#)

Latest blog posts

- 04 Sep 2013: [Ensembl 73 has been released!](#)
- 26 Aug 2013: [Automatic gene annotation workshop in Cold Spring Harbor, 29-30th October](#)
- 21 Aug 2013: [New Human Assembly Coming](#)

[Go to Ensembl blog →](#)

Browse

The Ensembl
database
eukaryotic
information

Popular



Zebrafish

Zv9

★ [Log in to customize this list](#)

2. Go to BioMart

e!Ensembl ASIA BLAST/BLAT | **BioMart** | Tools | More ▼ Human (GRCh37) ▼ Login/Register

Current selection:
< all Species
Only searching Human

Only searching Human ▼ **PDK1**
61 results match **PDK1** when restricted to species: Human ✕

Restrict category to:

Gene	5
Transcript	21
Variation	3
Somatic Mutation	28
GeneTree	1
Protein Family	3

Per page:
10 25 50 100
Show all results in one page

Layout:
Standard Table

1. COPY [Did you mean... ▼](#)

PDK1 (Human Gene)
ENSG00000152256 2:173420101-173489823:1
pyruvate dehydrogenase kinase, isozyme 1 [Source:HGNC Symbol;Acc:8809]
[Variation table](#) • [Location](#) • [Regulation](#) • [Orthologues](#) • [Gene tree](#)

PDK1-006 (Human Transcript)
ENST00000431718 2:173420830-173435548:1
pyruvate dehydrogenase kinase, isozyme 1 [Source:HGNC Symbol;Acc:8809]
[Location](#) • [cDNA seq.](#) • [Protein seq.](#) • [Population](#) • [Protein](#)

PDK1-005 (Human Transcript)
ENST00000436490 2:173420355-173429415:1
pyruvate dehydrogenase kinase, isozyme 1 [Source:HGNC Symbol;Acc:8809]
[Location](#) • [cDNA seq.](#) • [Protein seq.](#) • [Population](#) • [Protein](#)

PDK1-007 (Human Transcript)
ENST00000439519 2:173421293-173429801:1
pyruvate dehydrogenase kinase, isozyme 1 [Source:HGNC Symbol;Acc:8809]
[Location](#) • [cDNA seq.](#) • [Protein seq.](#) • [Population](#) • [Protein](#)

PDK1-004 (Human Transcript)

ENSG00000152256

ENST00000431718

[New](#) [Count](#) [Results](#)[★ URL](#)[XML](#)[Perl](#)[Help](#)**Dataset**Homo sapiens genes
(GRCh37.p12)**Filters**

[None selected]

Attributes

Ensembl Gene ID

Ensembl Transcript ID

Dataset

[None Selected]

Ensembl Genes **73** ▾

Homo sapiens genes (GRCh37.p12) ▾

[New](#) [Count](#) [Results](#)[★ URL](#)[XML](#)[Perl](#)[Help](#)**Dataset**Homo sapiens genes
(GRCh37.p12)**Filters**Ensembl Gene ID(s) [e.g.
ENSG00000139618]: [ID-list
specified]**Attributes**Ensembl Gene ID
Ensembl Transcript ID**Dataset**

[None Selected]

Please restrict your query using criteria below

☐ REGION:☐ GENE:☐ TRANSCRIPT EVENT:☐ GENE ONTOLOGY:☐ EXPRESSION:☐ MULTI SPECIES COMPARISONS:☐ PROTEIN DOMAINS:☐ VARIATION:

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

Dataset

Homo sapiens genes
(GRCh37.p11)

Filters

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

Attributes

Ensembl Gene ID
Ensembl Transcript ID

Please restrict your query using criteria below

▣ REGION:

▣ GENE:

☐ Limit to genes ...

with ArrayExpress ID(s)

☒ Only

☐ Excluded

☒ ID list limit [Max 500 advised]

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

ENSG00000152256

ENSG00000152256

New Count Results

URL XML Perl Help

Dataset

Homo sapiens genes
(GRCh37.p11)

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

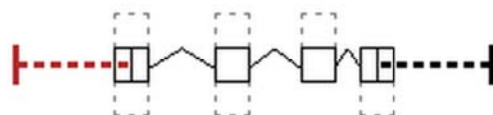
[None Selected]

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

- ☐ Features
- ☐ Structures
- ☐ Transcript Event
- ☐ Homologs
- ☐ Variation
- ☒ **Sequences**

☒ SEQUENCES:

Sequences (max 1)



- ☐ Unspliced (Transcript)
- ☐ Unspliced (Gene)
- ☐ Flank (Transcript)
- ☐ Flank (Gene)
- ☐ Flank-coding region (Transcript)
- ☒ **Flank-coding region (Gene)**
- ☐ 5' UTR
- ☐ 3' UTR
- ☐ Exon sequences
- ☐ cDNA sequences
- ☐ Coding sequence
- ☐ Protein

Upstream flank

☒ Upstream flank 2000

Downstream flank

☐ Downstream flank

 New  **Count**  Results

 URL  XML  Perl  Help

[Get result count for the query with any filters applied]

Dataset 1 / 63253 Genes

Homo sapiens genes
(GRCh37.p11)

Filters

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

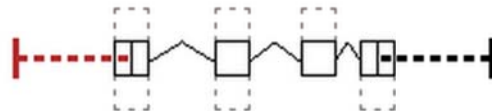
Attributes

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

- ☐ Features
- ☐ Structures
- ☐ Transcript Event
- ☐ Homologs
- ☐ Variation
- ☒ Sequences

SEQUENCES:

Sequences (max 1)



- ☐ Unspliced (Transcript)
- ☐ Unspliced (Gene)
- ☐ Flank (Transcript)
- ☐ Flank (Gene)
- ☐ 5' UTR
- ☐ 3' UTR
- ☐ Exon sequences
- ☐ cDNA sequences



10 min

New Count Results

URL XML Perl Help

Dataset 1 / 63253 Genes

Homo sapiens genes
(GRCh37.p11)

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]

Export all results to ☐ Unique results
only

Email notification to

View rows as ☐ Unique results only

```
>ENSG00000152256|ENST00000410055;ENST00000543905;ENST00000282077;E
GCAGCAAAATTGCAAACAGGGATTGAGGGAGCTGAATAATGCTGTGCATGATAAAGTGAC
AAGACTTTGGAGCTCAGGGAATAATCTAGAAGGAAGGAGATTAGAAAGGAGGAGGCAACC
TGAGAGGCTGGGGTAGAGCTGTGGAAATGTTTGAGGGACTATTTTAAGAATGTGTAATAT
TCAACTTAAAATTGTTTTCTTCTTTCCAACCTTAGGTCCCGCAGAAGAAGGGTGGGAAGAA
GAAGGGGGGCCATTTCAGCCATCAATGAGGTGGTGACACCATCAACATTCCCAAGCACATCC
CATTGAAACGGACTTCAGGAAGTGTGCCCCCTCGGGCACTCAGCAAAGATATCTGGAAATT
TGCCAAGAAAGAGATGGGCACTCCAGATGCGCATACTGATACCAGGTTTGACAAAGCTGT
CTGGCCAAAAGAATAATGTCCCATACGGTATCCACATGTGGTTGTCCAGAAAAACATAAT
GCGAATGAAGATTCACTGATATGTGAAAGGAAAATATCACGGGCCACAAAATCACTGAGC
```

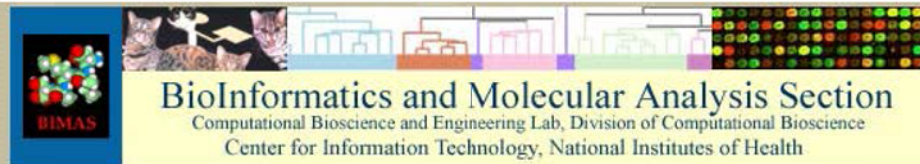
Web-based tools used today

- ENSEMBL & BioMart
 - Gene ID & sequence retrieval
- **Signal Scan**
 - Position-Specific Weight Matrix-based (PSWM)
 - <http://www.bimas.cit.nih.gov/molbio/signal/>
- **TFSEARCH**
 - PSWM-based
 - <http://www.cbrc.jp/research/db/TFSEARCH.html>
- **PROMO**
 - PSWM-based
 - http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
- The BEST
 - HMM-based

N-to-1
Multiple TFs to
ONE gene

Signal Scan

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



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

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):

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

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

The TRANSFAC database is Version 2.5 dated December 1995.

TRANSFAC [Signal Database](#) source

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):

```
>ENSG00000152256
GCAGCAAAATTGCAAACAGGGATTGAGGGAGCTGAATAATGCTGTGCATGATAAA
GTGACAAGACTTTGGAGCTCAGGGAATAATCTAGAAGGAAGGAGATTAGAAAGGA
GGAGGCAACCTGAGAGGCTGGGGTAGAGCTGTGGAAATGTTTGAGGGACTATTTT
AAGAATGTGTAATATTCAACTTAAAATTGTTTTCTTCTTTCCAACTTAGGTCCCG
CAGAAGAAGGGTGGGAAGAAGAAGGGGGCCATTTCAGCCATCAATGAGGTGGTGAC
```

**COPY &
PASTE**

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

submit

reset

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

Factor or Site Name		Loc.(Str.)		Signal Sequence	SITE #
3'_enhancer,_CD	site	957	(+)	TCTAAAAC	R02753
3'_enhancer,_HM	site	384	(-)	CATCTG	R04166
3'_enhancer,_HM	site	619	(-)	CATCTG	R04166
IRF-2	fact	54	(+)	AAGTGA	R00917
LF-A1	fact	1551	(+)	GGGCA	R01171
LF-A1	fact	1916	(-)	GGGCA	R01171
LF-A1	fact	325	(-)	GGGCA	R01171
LF-A1	fact	333	(+)	GGGCA	R01171
LF-A1	fact	376	(+)	GGGCA	R01171
LF-A1	fact	822	(+)	GGGCA	R01171
LF-A1	fact	859	(+)	GGGCA	R01171
MAZ	fact	1547	(+)	GGGAGGG	R02306
MEF-2	fact	159	(+)	CTATTTTAAG	R04212
MEP-1	fact	1499	(-)	TGCACAC	R03449
muEBP-C2	fact	453	(-)	CATGTG	R00886
muEBP-C2	fact	455	(+)	CATGTG	R00886
NF-1	fact	1078	(-)	CCAAT	R01543
NF-1	fact	1544	(-)	CCAAT	R01543
NF-1	fact	753	(+)	CCAAT	R01543
NF-1	fact	1022	(+)	GCCA	R01681

Alternative View on map

TRANSFAC ▾ [Signal Database](#) source

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

[illegible]

TFSEARCH

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



TFSEARCH: Searching Transcription Factor Binding Sites (ver 1.3)

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

Since 01/Apr/98 03630746



[Help](#)

Enter a label for the sequence:

Enter your DNA sequence below (by copy & paste):

Maximum Length : 9,999 b

Enter a label for the sequence:

PDK1 2K Promoter

Enter your DNA sequence below (by copy & paste):

Maximum Length : 9,999 b

```
>ENSG00000152256
GCAGCAAAATTGCAAACAGGGATTGAGGGAGCTGAATAATGCTGTGCATGATAAAGTGAC
AAGACTTTGGAGCTCAGGGAATAATCTAGAAGGAAGGAGATTAGAAAGGAGGAGGCAACC
TGAGAGGCTGGGGTAGAGCTGTGGAAATGTTTGAGGGACTATTTTAAGAATGTGTAATAT
TCAACTTAAATTGTTTTCTTTCTTTCCAACCTTAGGTCCCGCAGAAGAAGGGTGGGAAGAA
GAAGGGGGCCATTCAGCCATCAATGAGGTGGTGACACCATCAACATTCCTCAAGCACATCC
CATTGAAACGGACTTCAGGAAGTGTGCCCCCTCGGGCACTCAGCAAAGATATCTGGAAATT
TGCCAAGAAAGAGATGGGCACTCCAGATGCGCATACTGATACCAGGTTTGACAAAGCTGT
CTGGCCAAAAGAATAATGTCCCATACGGTATCCACATGTGGTTGTCCAGAAAAACATAAT
GCGAATGAAGATTCAGTATGTGAAAGGAAAATATCACGGGCCACAAAATCACTGAGC
TAAAGGGAAAATTCAAGCTGGGAACTGCTCAAGGCAAACCTGCCTCCCATTCTATTCAA
GTCATCCCTCTGCTGAGACAGATGCATATCTGATTGCTTCCTTGGGAAAGGCTTATCAGA
AACTCCAAAGAGCTGCAACCATTTGCTCTCTCAGCTAGTTATTAACCTGCAAGGCTTCCTGGC
```

Classification:

☐ use ALL matrices, ☒ Vertebrate, ☐ Arthropod, ☐ Plant, ☐ Yeast

Threshold score (default:85.0):

To execute the query, press this button:

Exec

To clear the form, press this button:

Clear

TFSEARCH Search Result

** TFSEARCH ver.1.3 ** (c)1995 Yutaka Akiyama (Kyoto Univ.)

This simple routine searches highly correlated sequence fragments
versus TFMATRIX transcription factor binding site profile database
by E.Wingender, R.Knueppel, P.Dietze, H.Karas (GBF-Braunschweig).

<Warning> Scoring scheme is so straightforward in this version.
score = 100.0 * ('weighted sum' - min) / (max - min)
The score does not properly reflect statistical significance!

Database: TRANSFAC MATRIX TABLE, Rel.3.3 06-01-1998
Query: PDK1 2K Promoter (2004 bases)
Taxonomy: Vertebrate
Threshold: 85.0 point

TFMATRIX entries with High-scoring:

	entry	score
1 ENSGGCAGCA AAATTGCAAA CAGGGATTGA GGGAGCTGAA TAATGCTGTG	M00101 CdxA	97.9
51 CATGATAAAG TGACAAGACT TTGGAGCTCA GGGAATAATC TAGAAGGAAG	M00147 HSF2	92.9
	M00240 Nkx-2.	90.7
	M00100 CdxA	88.5
	M00101 CdxA	86.4
	M00050 E2F	86.2

Database: TFFACTOR

Dated 1998

TRANSFAC database, binding factor information

Release 3.3, Jan 98

E.Wingender, R.Knueppel, P.Dietze, H.Karas (GBF-Braunschweig)
2,285 entries

PROMO

PROMO HOME PAGE

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 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:

TRANSFAC v8.3

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 the ALGGEN server: PROMO 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.

[SearchSites](#): input a query sequence to search for potential binding sites

or [MultiSearchSites](#): input a set of sequences to search for binding sites that different sequences share.

Other tools:

[ViewMatrices](#): view details on weight matrices being used.

練習

選擇一個你有興趣的基因，分析其 promoter 可否被 CREB 調控？並列出其 core 及 flanking sequence。

StAR
Steroidogenic Acute
Regulatory protein

Web-based tools for practice



10-15 min

- **ENSEMBL & BioMart**
 - Gene ID & sequence retrieval
- **Signal Scan**
 - Position-Specific Weight Matrix-based (PSWM)
 - <http://www.bimas.cit.nih.gov/molbio/signal/>
- **TFSEARCH**
 - PSWM-based
 - <http://www.cbrc.jp/research/db/TFSEARCH.html>
- **PROMO**
 - PSWM-based
 - http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
- **The BEST**
 - HMM-based

Only searching Human ▼

StAR



120 results match **StAR** when restricted to

species: Human ✕

[STAR \(Human Gene\)](#)

ENSG00000147465 8:38001167-38008783:-1

Steroidogenic acute regulatory protein [Source:HGNC Symbol;Acc:11359].

[Variation table](#) • [Location](#) • [Regulation](#) • [Orthologues](#) • [Gene tree](#)

ENSG00000147465

Search Promotor Sites

15 **Maximum matrix dissimilarity rate**

STRING TO BE ANALYZED

```
>ENSG00000147465
AGGCCTCGTGGGGCCTTGCTCAGCCTGACAGGGCCCCCTAAAGTGCAAAGACTAGGGTGAA
CACCCCAACGCCCTGTCTTTCTTTTTTTTCTTTTTTTTTTTTGGCAACAGAGTCTCGCG
CTCTGTCAATCCAGGCTGGAATGCAATGGTGTGATCTCAGATCACTGCAACCTCTACCTCC
CAAGTTCAAGCAATTCTGCCTCAGCCTCCTGAGTAGCTGGGATTACAGGTGCCTGCCACT
ATGCCTGGCTAATTTTTGTTTTTTTGTGTTTTGTTTTGTTTTGAGACGGAGTCTTGCTCT
GTCACCCAGGCTGGAGTACAGTGGCTTTATCTTGGCTCACTGCAACCTCTGCCTCCCGGG
TTCAAGTGATTATCCTGCCTCAGTCTCCTGAGTAGCTGGGATTATAGGCATACGCTACCA
CTCCCGGCTAATTTTTGTATTTGTCAGTAGAGACGAGGTTTACCATGTTGGTCAGGCTGG
```

or

FILE TO BE ANALYZED (in FASTA format)

Browse... No file selected.

Submit

0 C/EBPbeta [T00581]	1 c-Jun [T00131]	2 GR-alpha [T00337]	6 C/EBPalpha [T00108]	7 SRY [T00997]
8 TCF-4E [T02878]	9 GR [T05076]	10 VDR [T00885]	14 RXR-alpha [T01345]	15 PR A [T01661]
16 NFI/CTF [T00094]	17 NF-AT2 [T01945]	18 TFII-I [T00824]	22 c-Jun [T00132]	23 HNF-4alpha2 [T00132]
24 HNF-4alpha1 [T02429]	25 c-Jun [T00133]	26 PEA3 [T00685]	30 LXR-alpha:RXR-alpha [T05320]	31 VDR [T01369]
32 HNF-1A [T01211]	33 SREBP-1c [T01562]	34 LyF-1 [T00479]	38 ENKTF-1 [T00255]	39 HNF-3beta [T01005]
40 T3R-beta1 [T00851]	41 GATA-1 [T00305]	42 GATA-2 [T00308]	46 C/EBPalpha [T00105]	47 YY1 [T00915]
48 TFIID [T00820]	49 c-Myb [T00137]	50 ER-alpha [T00261]	54 ETF [T00270]	55 IRF-1 [T00423]
56 NF-AT1 [T00550]	57 c-Ets-2 [T00113]	58 c-Ets-1 [T00111]	62 MAZ [T00490]	63 WT1 [T00899]
64 AP-2beta [T02469]	65 AP-1 [T00032]	66 HNF-4alpha1 [T00372]	70 LEF-1 [T02905]	71 TBP [T00794]
72 MEF-2A [T01005]	73 HNF-3alpha [T02512]	74 HOXD8 [T01426]	78 PU.1 [T00702]	79 MyoD [T00526]
80 AR [T00040]	81 AP-1 [T00029]	82 EBF [T05427]	86 USF2 [T00878]	87 ATF3 [T01313]
88 NF-E2 [T00557]	89 USF-1 [T00875]	90 CAC-binding protein [T00076]	94 HNF-1A [T00368]	95 HNF-1B [T01950]
96 POU2F2 (Oct-2.1) [T00646]	97 RAR-beta:RXR-alpha [T05420]	98 RAR-alpha1 [T00719]	102 C/EBPgamma [T02028]	103 HOXD9 [T01424]
104 HOXD10 [T01425]	105 HOXD9 [T01755]	106 HOXD9 [T01756]	110 CREBbeta [T02361]	111 TCF-4 [T02918]
112 PU.1 [T02068]	113 HES-1 [T01649]	114 AhR:Arnt [T05394]	118 RelA [T00594]	119 c-Fos [T00124]

Zoom	Data (txt)															
	1	10	20	30	40	50	60	70	80	90	100	110	120	130	140	
ENSG00000147465		2 27	0 1 2 3 4 6 27	2 3 4 25 27	2 3 4 5	0 6 7 8 9 23 24 70	2 10 11 12 13 14 15 31 32 33 40	0 3 4 11 16 44	2	2 17 18 19 20 21 22 55 56 57 58 78 101	9 23 24	0 6 11 13 15 48		2 25 26 27 29	6 18 19 21 28 29 35	0 23 32 33 47

CTRL + F: CREB

110

CREBbeta [T02361]

was predicted in:

Sequence TGACGGAG

1155

1162

Dissimilarity 13.45%

RE equally 0.36621

RE query 0.37339

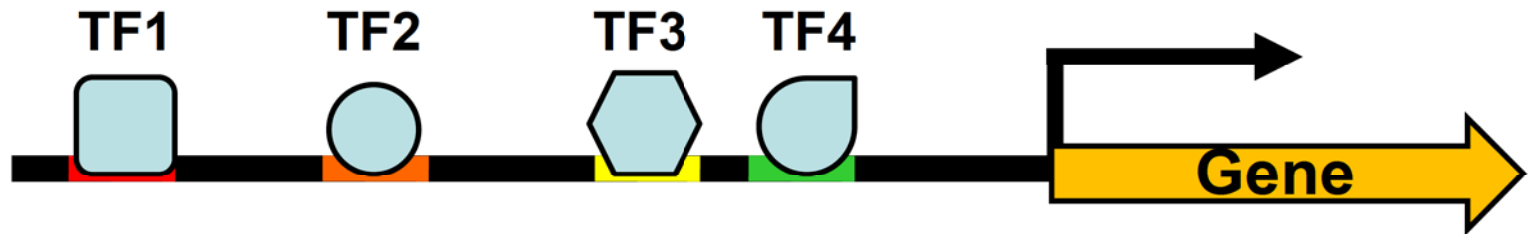
Consensus sequence and matrix:

A	0	0	2	0	0	0	0	3
C	0	0	0	3	0	0	3	0
G	0	2	1	0	3	0	0	0
T	2	1	0	0	0	3	0	0

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

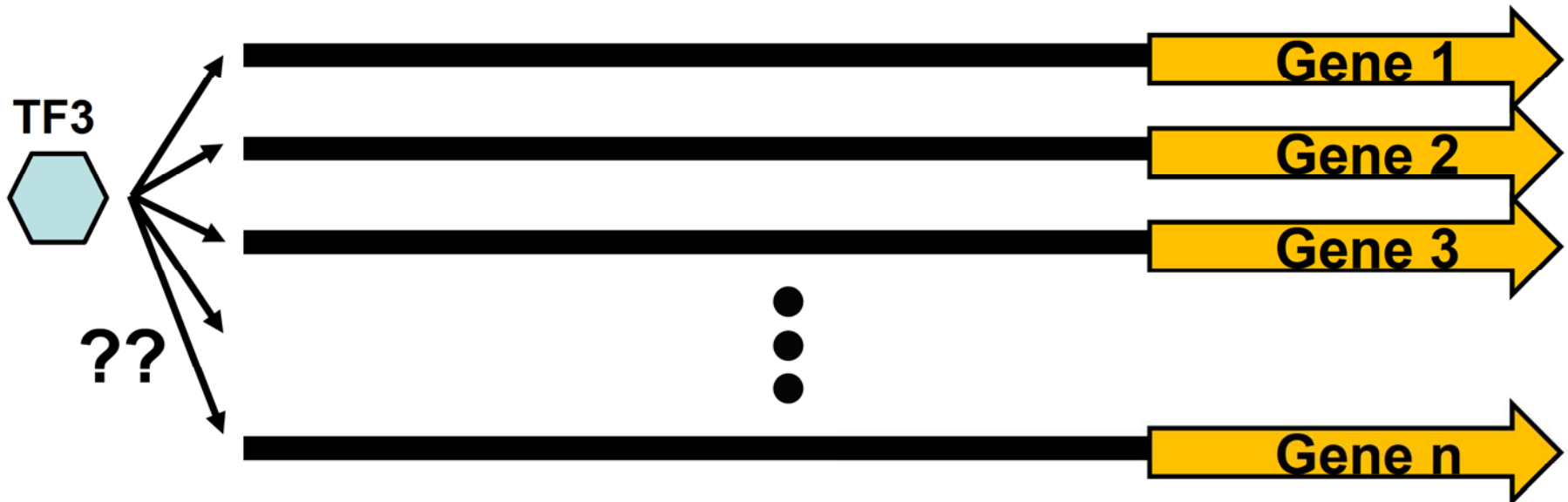
...TGCAACCACTGGGTATTTATTTA
 TTTATTTATTTAATT**TATTTATTT**
ATGACGGAG**TCTCACTCTG**TCGCC
 CAGGATGGAGTGCAGTGGCACGA...

ONE gene regulated by multiple transcription factors



How to predict genes targeted by single TF?

1-to-N
Single TF
regulates
multiple genes



Part III

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



Identifying transcription factor binding sites

- Consensus (string) search
- Position-Specific Scoring/Weight Matrix (PSSM or PSWM)
- Hidden Markov Model

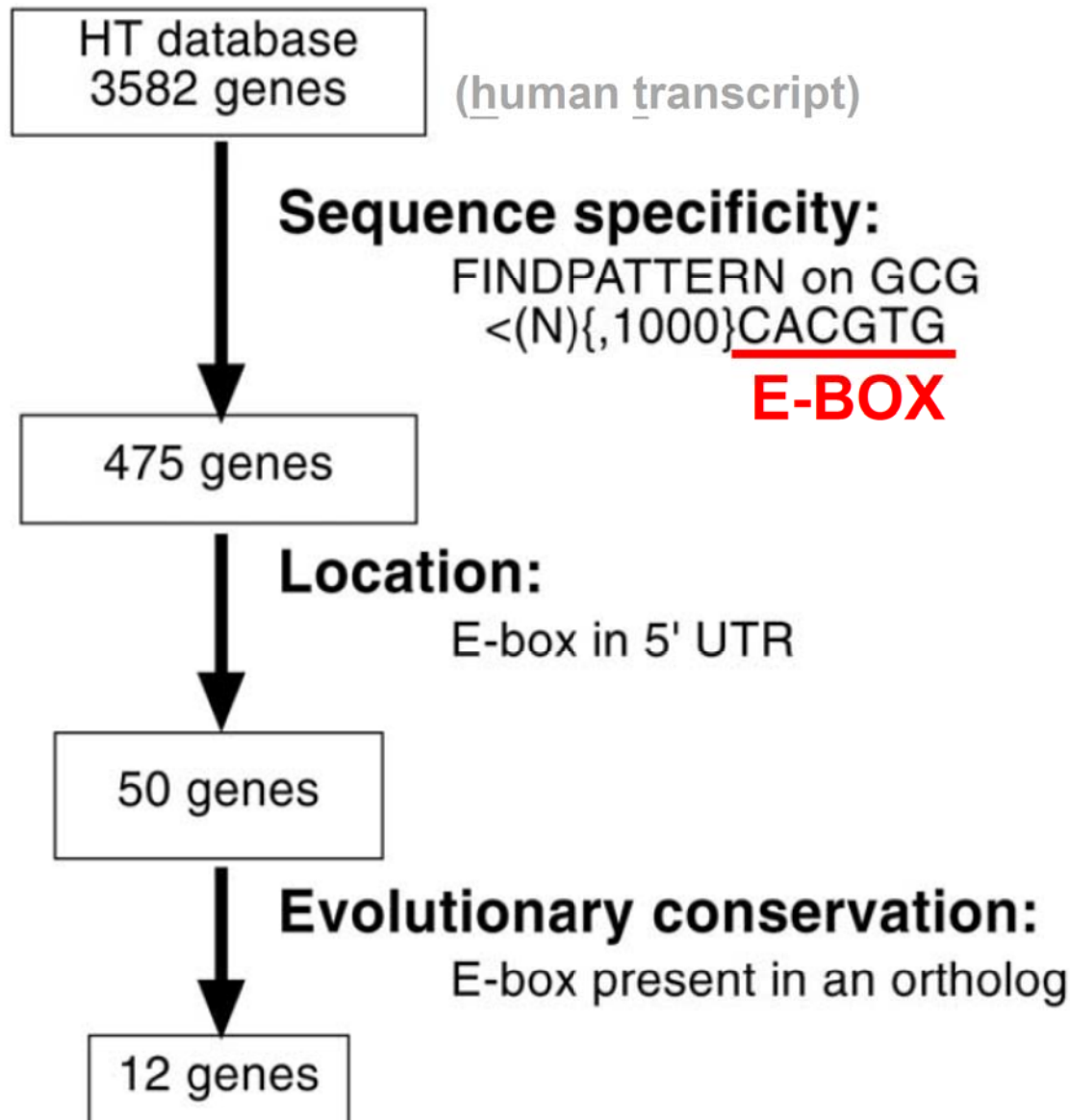
>StAR_2000

...TGCAACCACTGGGTATTTATTTA
 TTTATTTATTTAATT**TATTTATTT**
ATGACGGAGTCTCACTCTGTCGCC
 CAGGATGGAGTGCAGTGGCACGA...

CRE SEQ.

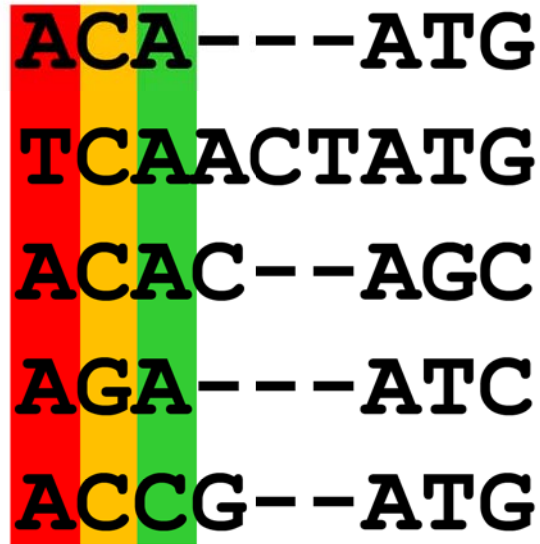
CON	:	TGACGTCA	0
SITE1:		TGAC CT CA	?
PRED	:	TGACGGAG	1
SITE2:		AGACGGAG	?

Consensus Search – c-MYC as example



PSSM (TRANSFAC database)

A multiple alignment-based motif profiling



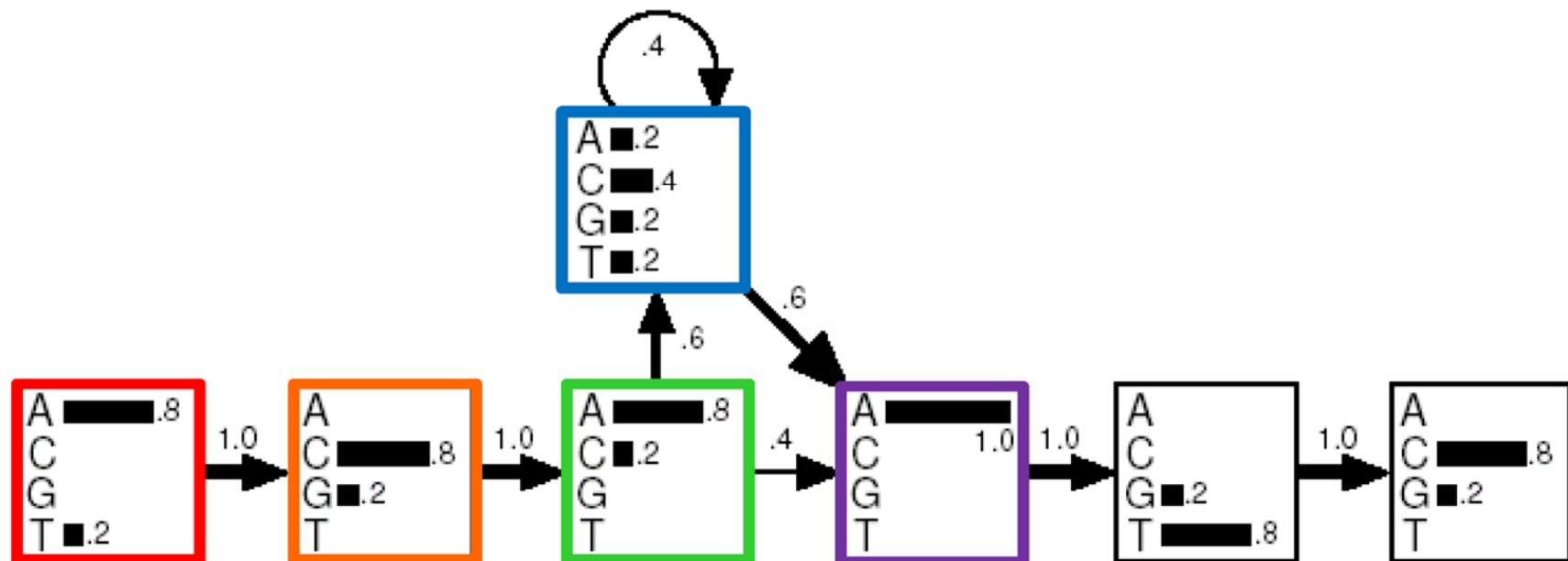
ACA---ATG
TCAACTATG
ACAC--AGC
AGA---ATC
ACCG--ATG

	1 ST	2 ND	3 RD	4 th
A	.8	0	.8	?
G	0	.2	0	?
C	0	.8	.2	?
T	.2	0	0	?

- (1) Gapping is a problem
- (2) Occurrence of individual base is NOT a independent event

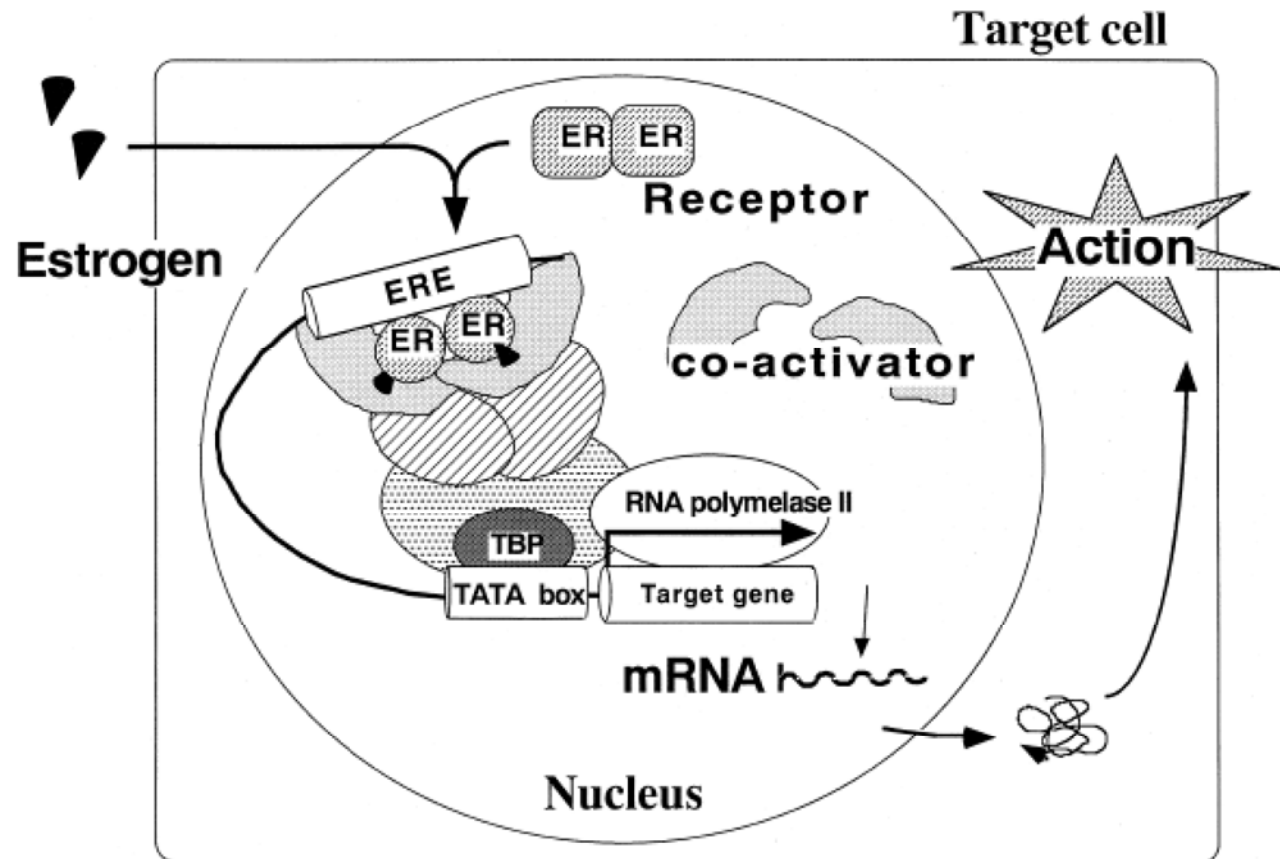
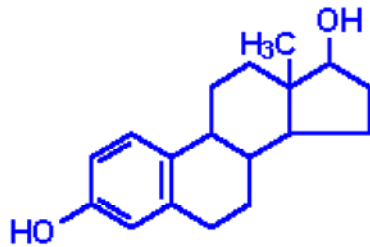
Hidden Markov Model

ACA---ATG
TCAACTATG
ACAC---AGC
AGA---ATC
ACCG---ATG



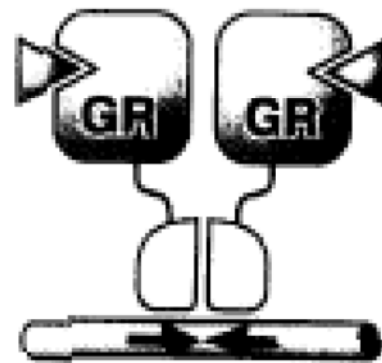
Pathway of the Estrogen Receptor (ER)

17 β -Estradiol



Estrogen Response Element (ERE)

Steroid Receptors



GR	<i>glucocorticoid</i>
MR	<i>mineralcorticoid</i>
PR	<i>progesterone</i>
AR	<i>androgen</i>
ER	<i>estrogen</i>

palindrome

GGTCANNNTGACC
CCAGTNNNACTGG

The Binding Element Searching Tool (The BEST)

- **Input:**
 - A set of seq. or multiple seq. alignment
- **HMMER builds hidden Markov model**
- **Search a local database containing mouse, rat and human promoter sequences with GO annotations**

Motivation

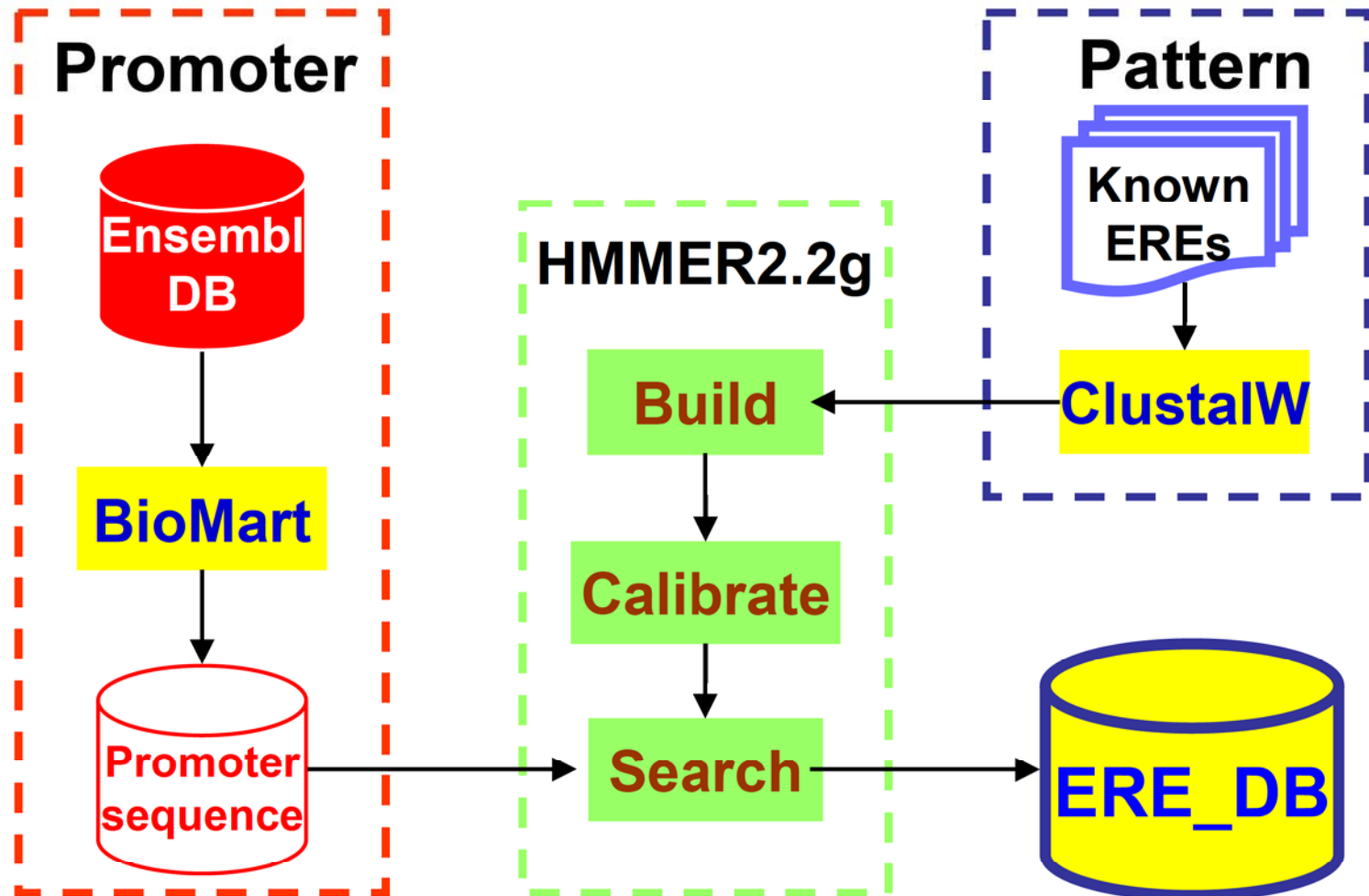
- Search for novel estrogen regulated genes by known ERE based on multiple seq. alignment



Literature search
(promoter activity assay,
gel shift assay or ChIP)

System modules

Hidden Markov Model based search tools:



◎應用資源

miRNA prediction

21)

生物資訊中心將於2013/8/27(二)進行網路換線工程，於2013/8/27(二)8:00停止各系統服務，2013/8/27(二)18:00開始恢復各系統運作，若有不便敬請見諒！

cility)，將於
基因體分析研習會，
基因體分析網路工
方式讓學員們快速
呈內容包括
名時間：

公告日期 2013-08-



NCKU Bioinformatics Center

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

[Introduction](#)[Start](#)[My Result](#)[ReadMe](#)

The **B**inding **E**lement **S**earching **T**ool is an online genome-wide searching tool.

Users can 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. Results can be further analyzed by gene function and expression pattern via EnsMart interior linking to Gene Ontology and linkout to AFFY database.



© By [NCKU Bioinformatics Center](#), R.O.C. Government 2004. All Rights Reserved.

THE Binding Element Searching Tool

Introduction

Start

My Result

ReadMe

STEP 1. Multiple Sequence Alignment	
a.	Enter a set of sequences in FASTA format
	<pre>>ENSG00000171105 TACACAAGTGAaacTGGCCCAGAGA >ENSG00000135744 TAAATAGGGCAtcgTGACCCGGCCG >ENSG00000147654 TTTCCGGGTCAgggTGACCTCTGGG</pre>
b.	I want using my pre-aligned seq. in MSF format: 選擇檔案 未選擇檔案
NEXT	

NCKU Bioinformatics

THE Binding Element

Introduction Start My Result

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
ENSG00000138061	CTGCCAGGTCGCGGTGCCCTCCTTC	
ENSG00000160182	CTGCAAGGTCACGGTGGCCACCCCG	
ENSG00000087510	ACGCCTGGTCACCGTGACCCCGATT	
ENSG00000115944	TGCAGGGGTCAAGGTGACCCCGGG	
ENSG00000085465	TTCTGGGGTCACTGTGACTCTCATA	
ENSG00000147654	TTTCCGGGTCAGGGTGACCTCTGGG	
ENSG00000117984	AAGCTGGGCCGGGOTGACCCGCGG	
ENSG00000164362	CATGTTGGTCAGGOTGATCTCAAAA	
ENSG00000135744	TAAATAGGGCATCGTGACCCGGCCG	
ENSG00000118777	GTGTCACGGCAGGGTGACCCTAGCC	
ENSG00000171105	TACACAAGTGAACTGGCCAGAGA	

Scores for complete sequences (score includes all domains):

Sequence	Description	Score	E-value	N
ENSG00000115944		9.2	0.0015	1
ENSG00000147654		7.5	0.0038	1
ENSG00000087510		7.1	0.0049	1
ENSG00000160182		6.5	0.0068	1
ENSG00000117984		5.6	0.011	1
ENSG00000135744		4.5	0.021	1
ENSG00000085465		3.8	0.030	1
ENSG00000164362		3.7	0.033	1
ENSG00000118777		3.6	0.034	1
ENSG00000138061		2.4	0.067	1
ENSG00000171105		-0.3	0.3	1

STEP 1. Multiple Sequence Alignment

Done!

View [emma result](#).

STEP 2. Construct and calibrate a HMM

Done!

View [hmmbuild result](#), [hmmcalibrate result](#), [hmmsearch with original sequences result](#).

STEP 3. Search the specific HMM in a dataset

sets T bit threshold (globT) to >= -0.3

☐ Human

a.

☐ Mouse

☐ Rat

Because this step needs some time to complete, please remember your account "uo1gu0s2po" for checking your result later. After clicking "NEXT" button, it's ok to close the browser.

Your mail :

BACK

NEXT

uo1gu0s2po

NCKU Bioinformatics Center

THE Binding Element Searching Tool

[Introduction](#) [Start](#) [My Result](#) [ReadMe](#)

	Enter your account:	
	NEXT	

uo1gu0s2po

© By [NCKU Bioinformatics Center](#) , R.O.C. Government 2004. All Rights Reserved.

NCKU Bioinformatics Center

THE Binding Element Searching Tool

☐ Introduction

☐ Start

☐ My Result

☐ ReadMe

	Select:
	☒ Human
	BACK NEXT



© By [NCKU Bioinformatics Center](#), R.O.C. Government 2004. All Rights Reserved.

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**

Raw data – Table view

NCKU Bioinformatics Center

THE Binding Element Searching Tool

Introduction

Start

My Result

ReadMe

BACK

Sequence		seq-f	seq-t	hit-seq	score	E-value	orientaiton
-----		-----	-----	-----	-----	-----	-----
ENST00000508455	ENSG00000197530	4977	4996	GGGTCAAGCTGACCCCGAgG	9.3	19	reverse
ENST00000512004	ENSG00000197530	4518	4537	GGGTCAAGCTGACCCCGAgG	9.3	19	reverse
ENST00000486088	ENSG00000241656	1275	1294	AGGTCAAGGTGACCTCGGgA	8.3	33	reverse
ENST00000408417	ENSG00000221344	8	27	CGGTCAACGGTGACCCCCAgG	8.2	37	reverse
ENST00000429920	ENSG00000198925	149	168	CGGTCAACGGTGACCCCCAgG	8.2	37	reverse
ENST00000446716	ENSG00000198925	1375	1394	CGGTCAACGGTGACCCCCAgG	8.2	37	reverse
ENST00000475339	ENSG00000198925	1865	1884	CGGTCAACGGTGACCCCCAgG	8.2	37	reverse
ENST00000357972	ENSG00000147041	1417	1436	AGGTCAAGGTGACCCCAaA	8.1	38	reverse
ENST00000456733	ENSG00000147041	1417	1436	AGGTCAAGGTGACCCCAaA	8.1	38	reverse
ENST00000433814	ENSG00000185619	3523	3542	GGGGCAGGGTGACCCCCAgC	8.0	40	reverse
ENST00000516383	ENSG00000252192	297	316	AGGTCACTGTGACCCCAaTg	8.0	40	reverse
ENST00000475730	ENSG00000215788	298	317	AGGTCAAGGTGAGCCCCAaG	7.9	44	reverse
ENST00000293725	ENSG00000188868	1746	1765	AGGTCAAGGTGAGCCCTGGgG	7.7	48	reverse
ENST00000318168	ENSG00000188868	1714	1733	AGGTCAAGGTGAGCCCTGGgG	7.7	48	reverse
ENST00000390572	ENSG00000211912	1528	1547	AGGTCAAGGTGAGCCCTGGgG	7.7	48	reverse
ENST00000390571	ENSG00000211911	4045	4064	AGGTCAAGGTGAGCCCTGGgG	7.7	48	reverse

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](#)

NCKU Bioinformatics Center

THE Binding Element Searching Tool

Introduction

Start

My Result

ReadMe

Human menu:

a. [View all](#) [GO Clustering](#)

b. Field searching

you can use "" or -

Gene Description:

GO Description:

StAR

c. ENSG searching

ENSG list:

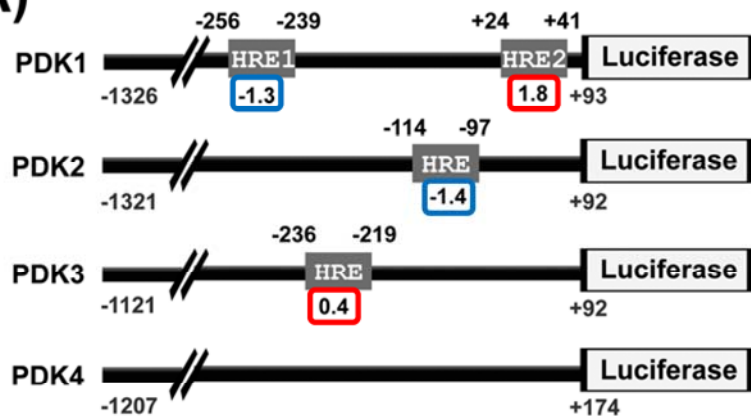

 BACK

1.

ENSG ID						
ENSG00000159433						
Gene						
STARD9 (DKFZp686O248, DKFZp686P19131, DKFZp781J069, FLJ16106, FLJ21936, KIAA1300)						
Gene Description						
StAR-related lipid transfer (START) domain containing 9 [Source:HGNC Symbol;Acc:19162]						
ENST	Score		Hit Seq.	Position		Orientation
ENST00000333475	0.8		AGTTCAGCCTGACCAACCTG	-811~-792		reverse
ENST00000333475	2.9		TGGTCAGGCTGAACTCGAAC	-806~-787		forward
ENST00000290607	0.8		AGTTCAGCCTGACCAACCTG	-868~-849		reverse
ENST00000290607	2.9		TGGTCAGGCTGAACTCGAAC	-863~-844		forward
AFFY ID						
227108_at						
3591076						
3591084						
GO Characterization						
Function	GO:0005524	ATP binding				
	GO:0003777	microtubule motor activity				
	GO:0000166	nucleotide binding				
Process	GO:0007018	microtubule-based movement				
InterPro Characterization						
IPR001752	Kinesin, motor domain					
IPR000253	Forkhead-associated (FHA) domain					

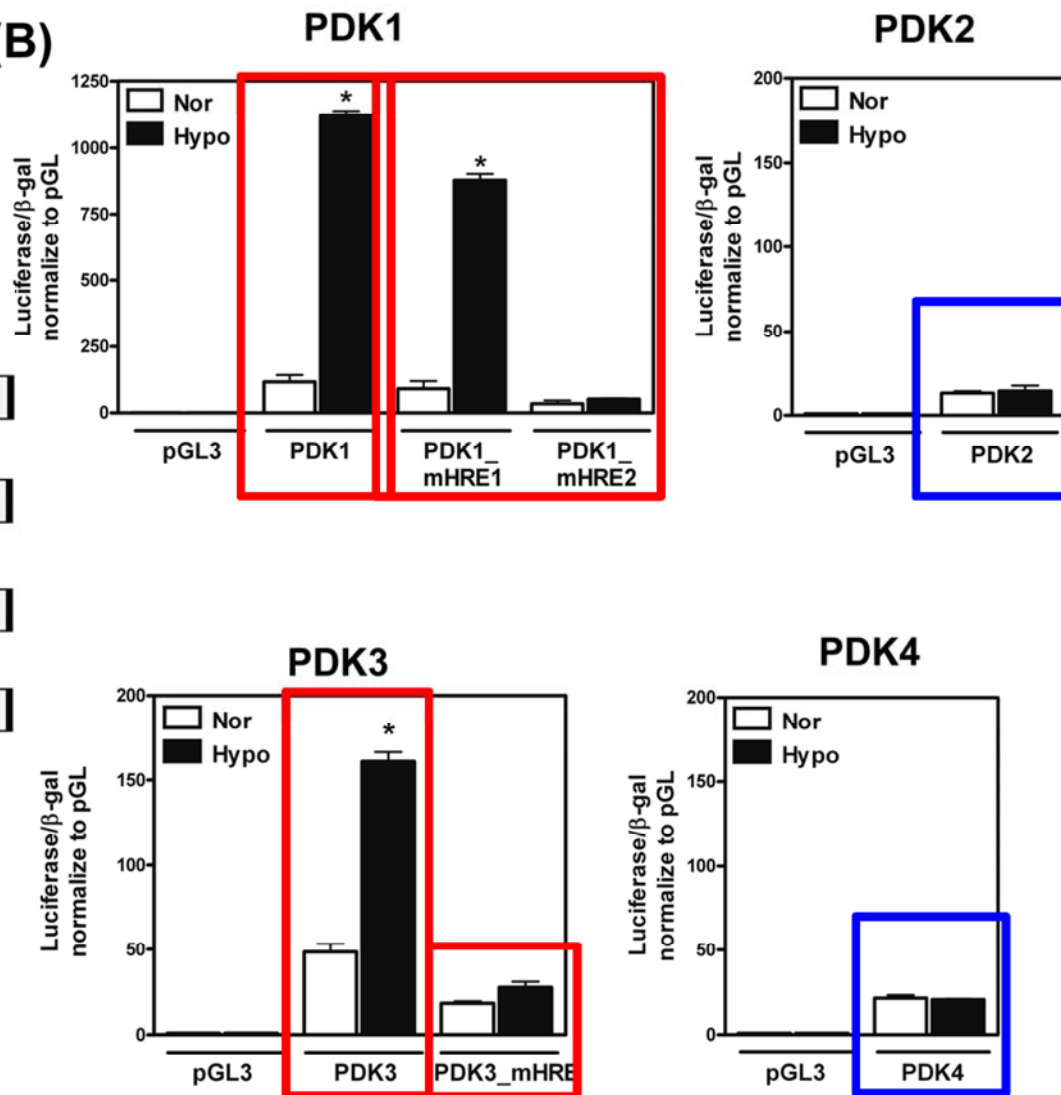
Hypoxia induces promoter activities of PDK1 and PDK3

(A)



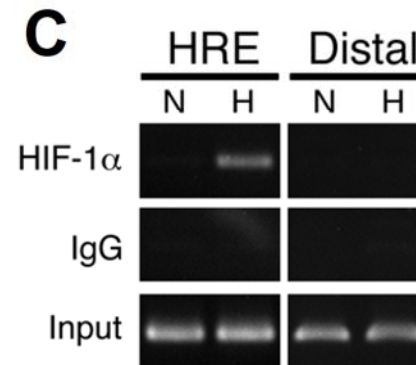
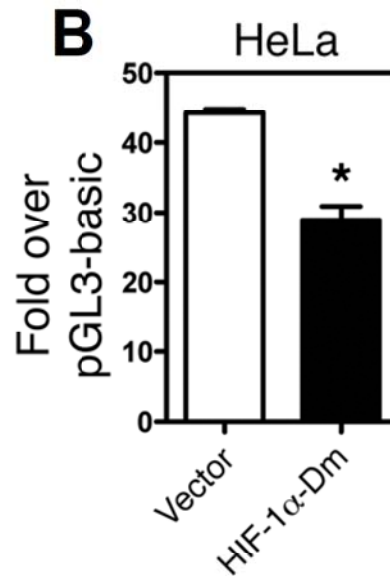
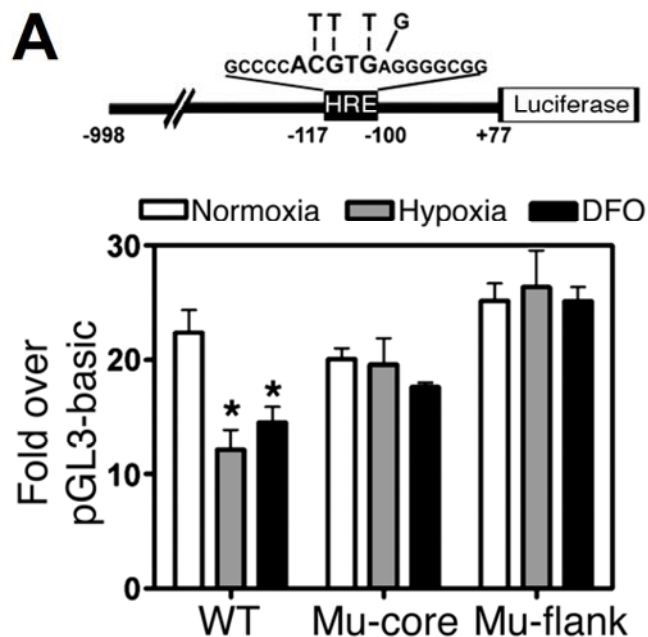
core HRE RCGTG
mHRE RAAAG

(B)



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



練習

- 利用 PubMed search 文獻，找出約 10 個 CRE (cAMP-response element) 的 sequences (需包含 5' 與 3' flanking sequences，全長約 20~25 bp 左右)，以 FASTA format 輸入 The BEST 進行分析，並列出 10 個 candidate genes。

Regulation of cAMP-inducible genes by CREB.

Montminy MR, Gonzalez GA, Yamamoto KK.

Clayton Foundation Laboratories for Peptide Biology, Salk Institute, La Jolla, CA 92037.

Abstract

A large number of neuropeptides and neurotransmitters stimulate neuronal cells through the second messenger cAMP. These synaptic signals often cause profound changes in neuronal function by altering basic patterns of gene expression. Cyclic AMP, in turn, regulates a number of these genes through a conserved cAMP response element (CRE). Recently, a nuclear CRE-binding protein, CREB, has been shown to bind to the CRE and stimulate the transcription of cAMP-responsive genes. This article reviews recent progress towards understanding the mechanism by which cAMP modulates the activity of CREB to stimulate gene transcription.

PMID: 1693237 [PubMed - indexed for MEDLINE]

TABLE I. The palindromic sequence, 5'-TGACGTCA-3', that is conserved among cAMP-responsive genes

Gene	Sequence
Somatostatin ^a	CTGGGGGCGCCTCCTTGGCT GACGTC AGAGAGAGAG (-32)
PEPCK ^a	TGATCCAAAGGCCGGCCCCCTT ACGTC AGAGGCGAGC (-74)
VIP ^a	TCCCATGGCCGTCATACTGT GACGTCTTT CAGAGCA (-60)
Parathyroid hormone	GGGAGT GACGTC ATCT (-65)
Proenkephalin ^a	GGGCCTGCGT CAGC (-87)
α-Chorionic gonadotropin	AAAATT GACGTC ATGG (-113)
c-fos ^a	CCGCC CAGTGACGT AGGA (-57)
Cytomegalovirus enhancer	CCACCCCAT TGACGTCA ATGGGAGTT (-124)
BLV LTR	ACCAGACAGAG ACGTCAGCTG CCAGA (-144)
HTLV-II LTR	CCACGGCCCT GACGTCCCT CCCCCCC (-162)
Intracisternal A particle	CCTCTCCCGT GACGTC ATCTGGGG (-86)

^a Genes known to be transcriptionally regulated by cAMP.
Boldface type indicates palindromic sequences.
Position of 3'-most nucleotide is indicated in parentheses next to each sequence.
Abbreviations: PEPCK, phosphoenolpyruvate carboxykinase; VIP, vasoactive intestinal peptide; BLV, bovine leukemia virus; LTR, long terminal repeat.

NCKU Bioinformatics Center

THE Binding Element Searching Tool

Introduction

Start

My Result

ReadMe

Enter your account:

NEXT

uo1gu0s2po

© By [NCKU Bioinformatics Center](#), R.O.C. Government 2004. All Rights Reserved.