# Understanding

## Promoter Analysis

Thomas Werner
CEO&CSO
Genomatix Software GmbH
Landsberger Strasse 6, D-80339 München
http://www.genomatix.de


© 2004 by Genomatix Software GmbH

# Double-check your expectations!

## What you will hear about...

### Part I : Understanding promoter elements

- Some basics about transcription control in general
- Properties of transcription factor binding sites and how to analyze them
- Where and how to obtain promoter sequences

### Part II : Understanding promoter function

- The difference between binding sites and transcriptional modules
- The difference between a physical promoter and promoter functions
- Comparative genomics with promoters

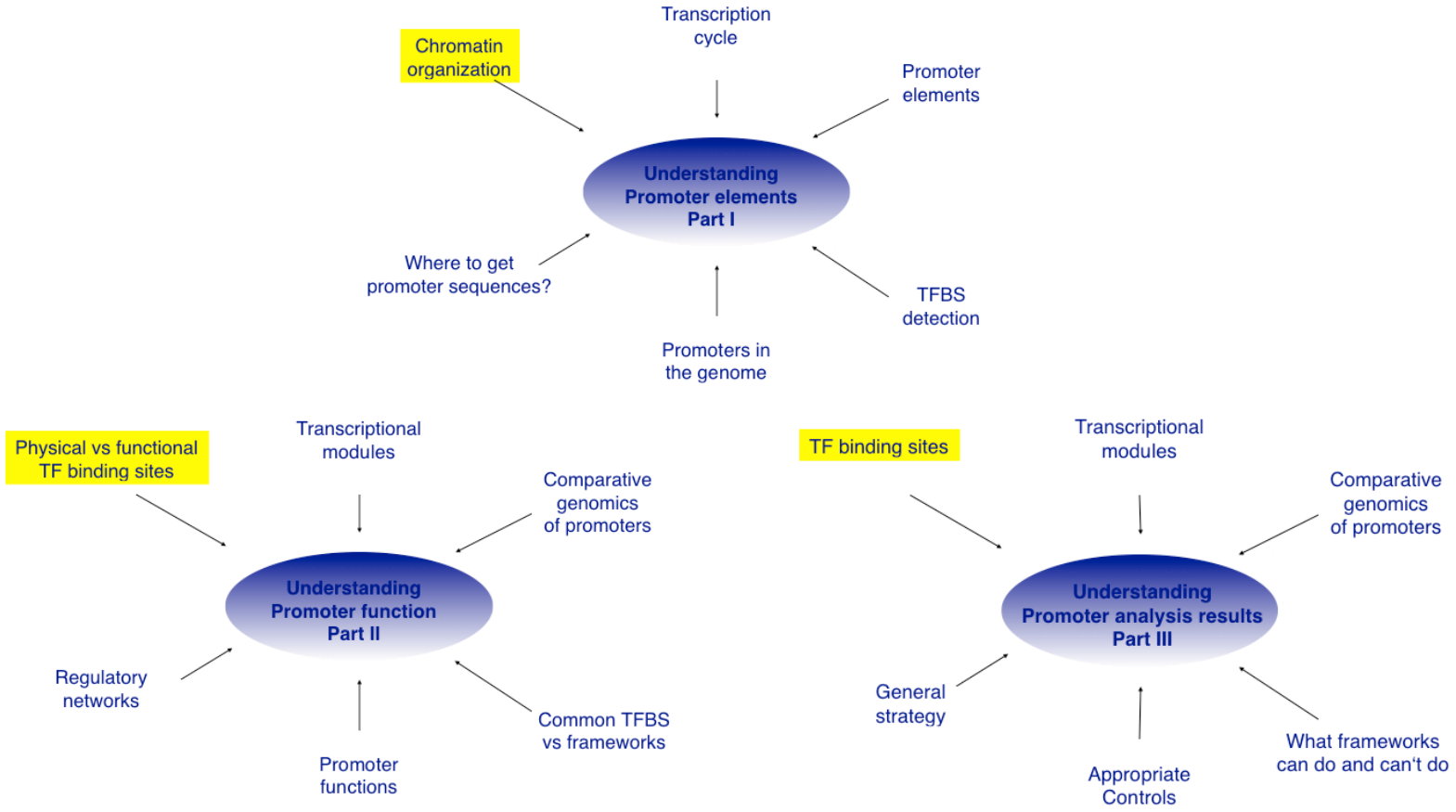
### Part III : Understanding promoter analysis results

- Strategies and tools to analyze and detect biological functionality
- You will see how to set up strategies with existing tools
  NOT how to generate new tools yourself


© 2004 by Genomatix Software GmbH

# Double-check your expectations!

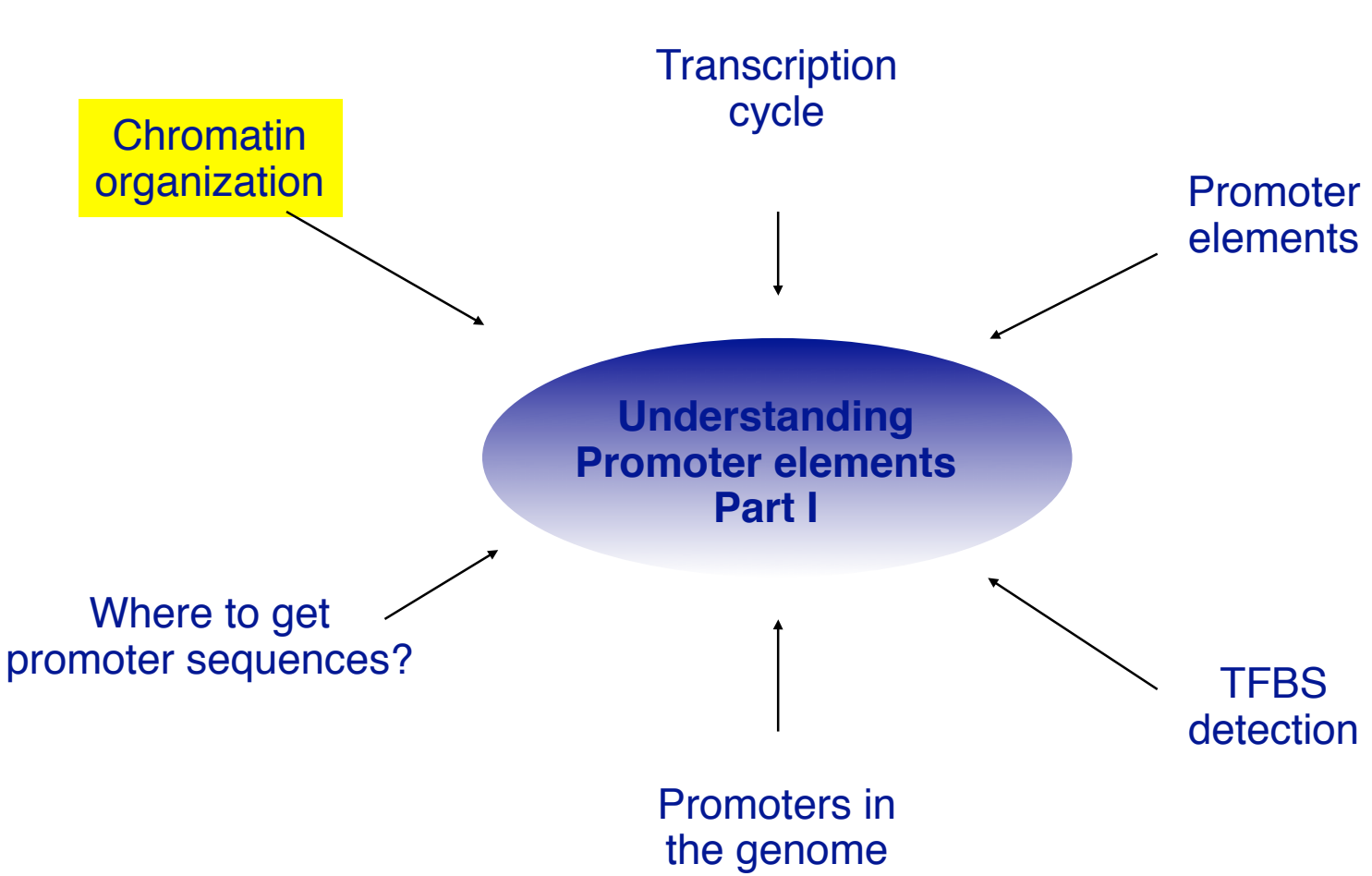
## Three times around the clock...





© 2004 by Genomatix Software GmbH

# Biological context of transcription



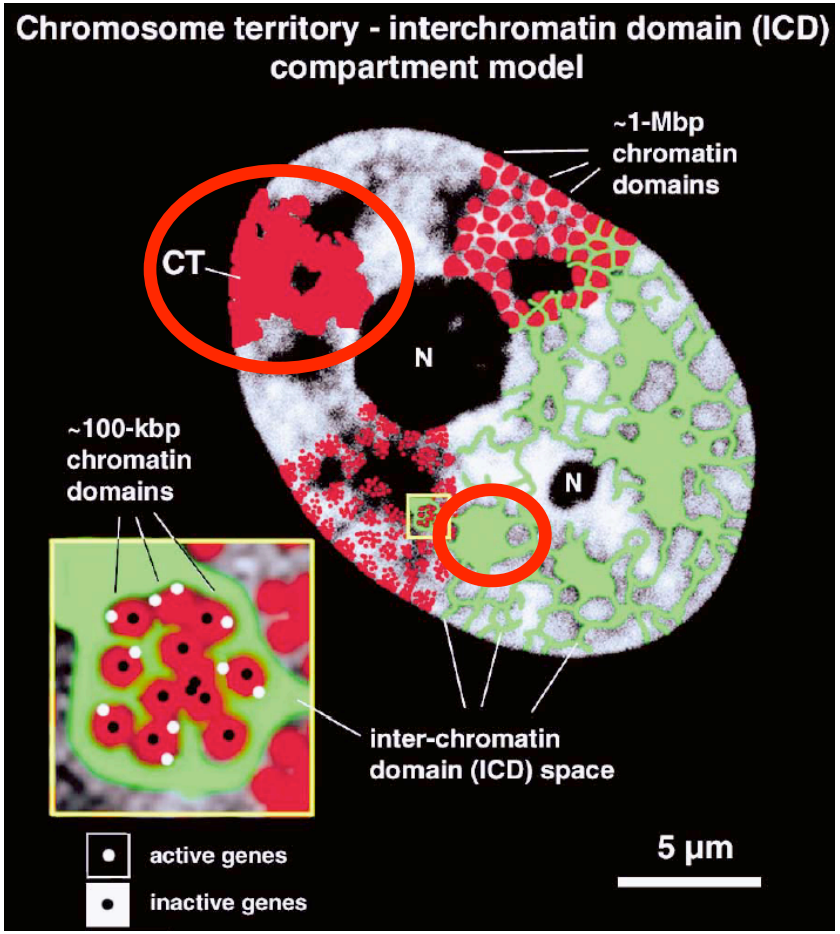


© 2004 by Genomatix Software GmbH

# Part I : Chromatin organization

## The chromatin in the nucleus is highly organized

- There are two different environments within a Cell nucleus
- Chromatin territories (transcriptionally inactive)
- Interchromatin domains (transcriptionally active)
- Chromatin moves between those compartments **(chromatin hypothesis T. Cremer)**

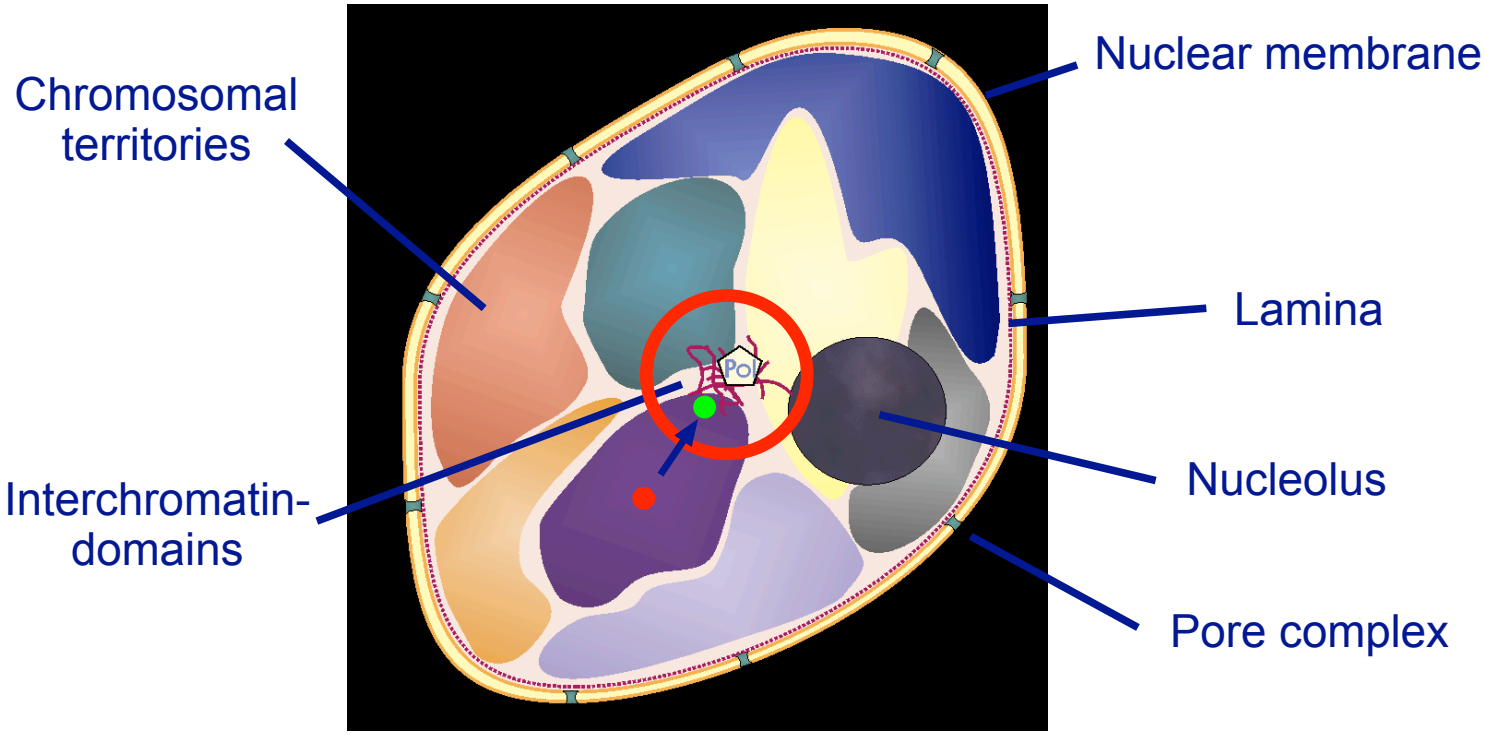




© 2004 by Genomatix Software GmbH

# Part I : Chromatin organization

## The chromatin in the nucleus is highly organized

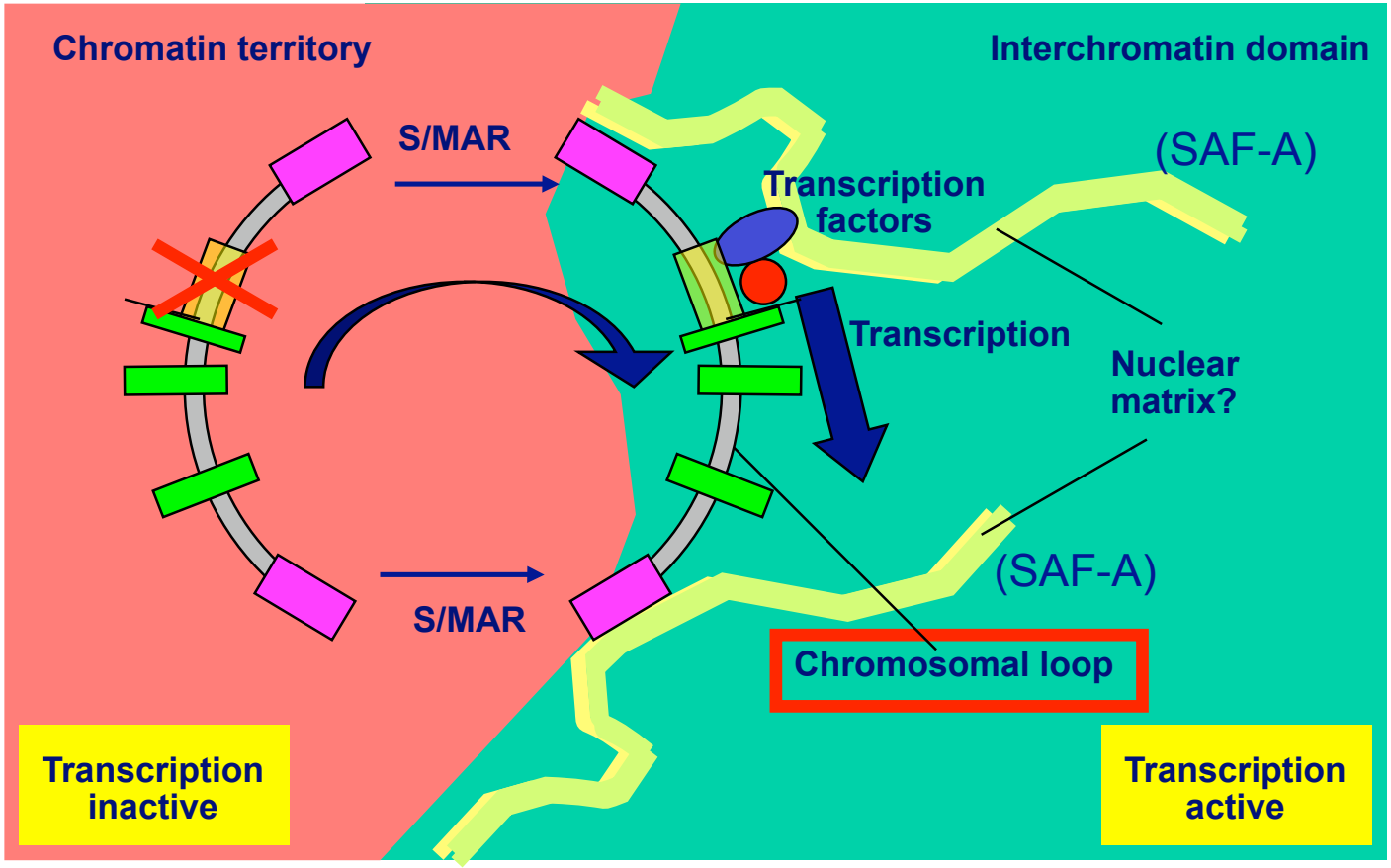



Inter chromatin domains are where the *in situ* nuclear matrix is located


© 2004 by Genomatix Software GmbH

# Part I : Chromatin organization

## Potential activation mechanisms of a chromosomal loop

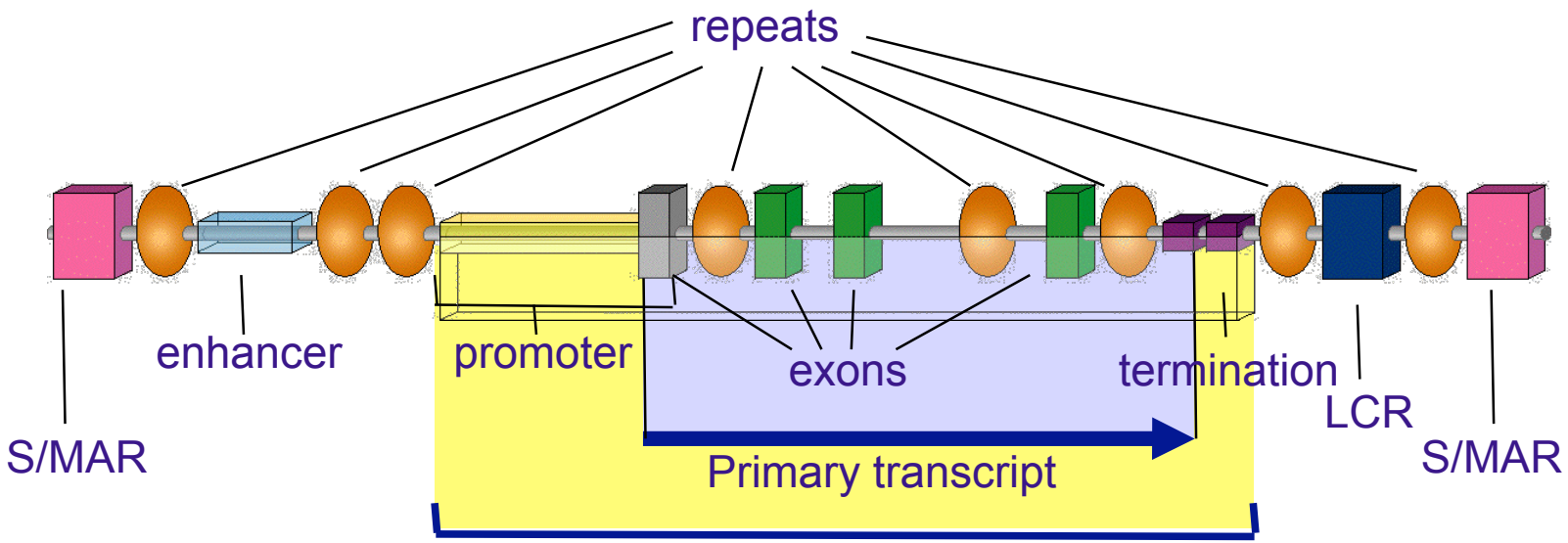




© 2004 by Genomatix Software GmbH

# Part I : Chromatin organization

## Schematic organization of a chromosomal loop



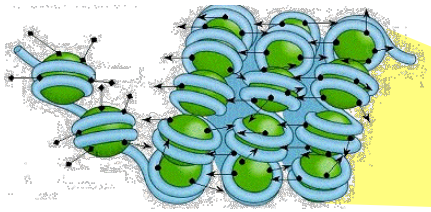

**(by definition includes all gene-specific regulatory sequences)**

Only part of the functional context is located in cis on the genomic DNA


© 2004 by Genomatix Software GmbH

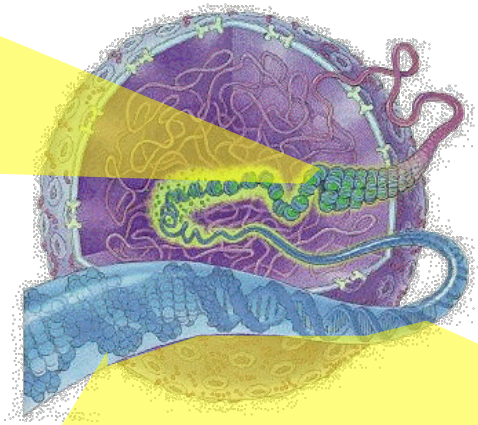
# Part I : Chromatin organization

## Tissue/cell specificity is determined by multiple events



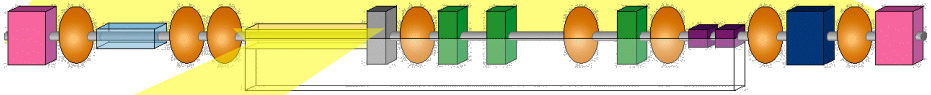
### Epigenetics

**Chromatin structure**
**DNA methylation**
**Histone de/acetylation**

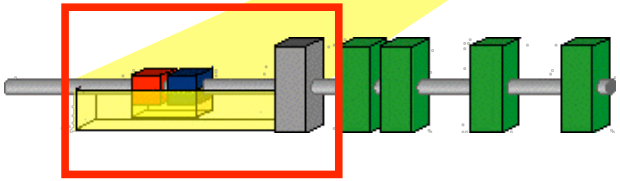


### Differentiation

**Permanent Heterochromatin**
**DNA recombination**



**Promoter analysis**



### Transcription control

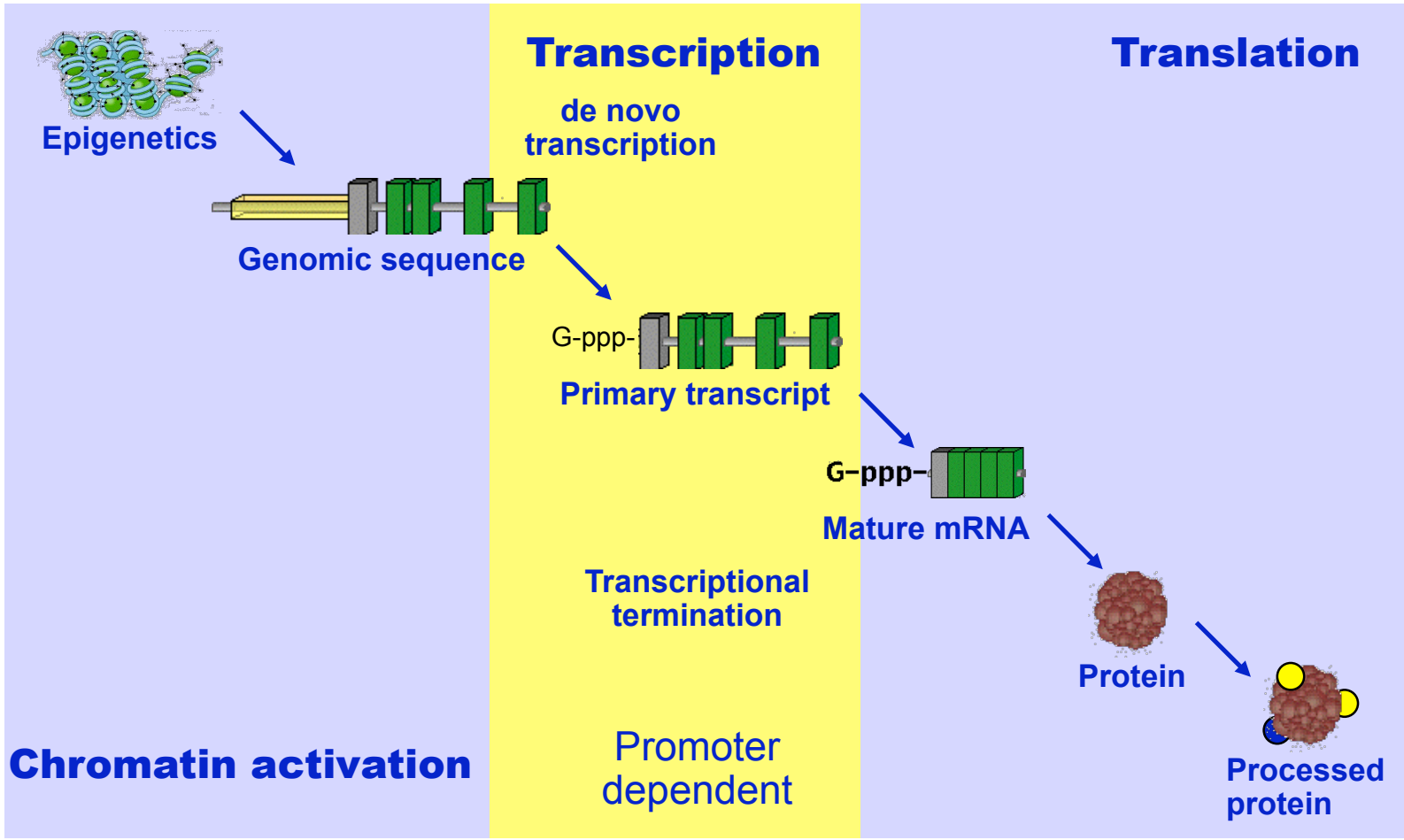
**Locus Control Regions**
**Enhancers**
**Promoters**



© 2004 by Genomatix Software GmbH

# Part I : Chromatin organization

## Which part of gene expression is controlled by promoters?






© 2004 by Genomatix Software GmbH

# Part I : Chromatin organization

## Summary

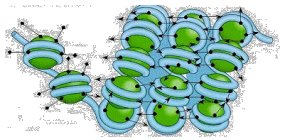
Gene expression is controlled on many levels

- Chromatin is transcriptionally inactive by default
- Activation includes relocation and S/MAR association
- Epigenetic control decides on accessibility of promoters
- Promoters act in transcriptional initiation

Promoters are the final processors in transcription control


© 2004 by Genomatix Software GmbH

# Biological context of transcription



Chromatin organization

Transcription cycle

Promoter elements

**Understanding Promoter elements Part I**

Where to get promoter sequences?

Promoters in the genome

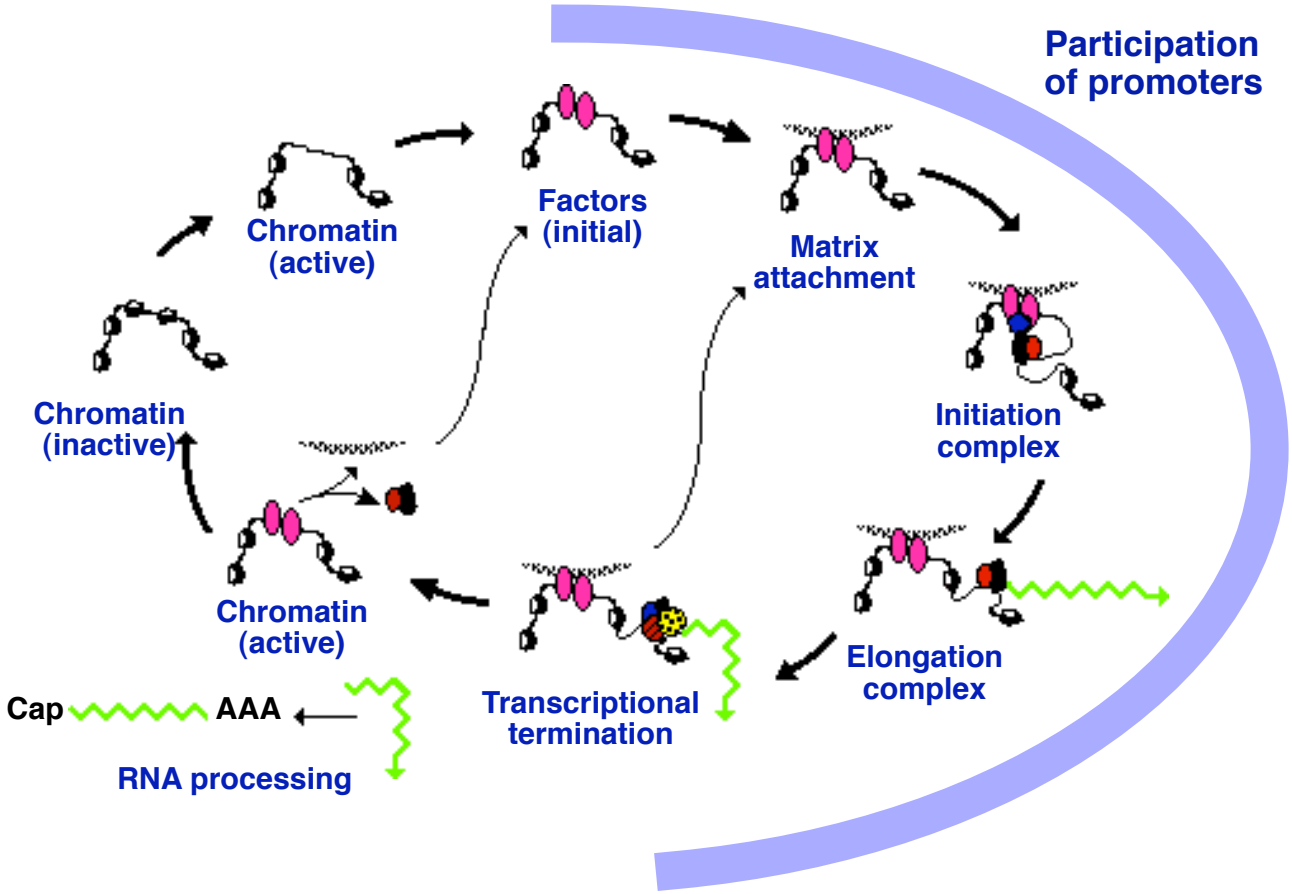
TFBS detection



© 2004 by Genomatix Software GmbH

# Part I : Transcription cycle

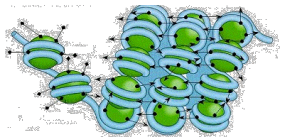
## Transcription cycle of chromatin





© 2004 by Genomatix Software GmbH

# Biological context of transcription



Chromatin organization

Transcription cycle

Promoter elements

**Understanding Promoter elements Part I**

Where to get promoter sequences?

Promoters in the genome

TFBS detection

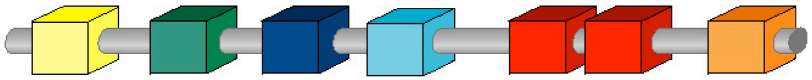


© 2004 by Genomatix Software GmbH

# Part I: Promoter elements

## Promoters contain discernible elements

- Promoter elements are necessarily in linear order within the promoter



- Very important elements are transcription factor binding sites (TFBSs)
- Promoters always contain at least one transcription start site (TSS)
- Promoters may also contain repeats and inverted repeats
- Promoters apparently also contain nucleosomal positioning elements

Promoters do not contain any known terminal elements


© 2004 by Genomatix Software GmbH

# Part I: Promoter elements

## Basic types of known promoter elements

- Transcription factor binding sites (TFBSs)

  Example: TGASTCA binds to AP-1

- Secondary structures (cruciform DNA, hairpins)

  Example: AAAAAGCTGTTTTT can form a hairpin

- Repeated sequences

  Example: TCCAGTAGTCCAGTGTCCAGT

These elements can be detected in the nucleotide sequence

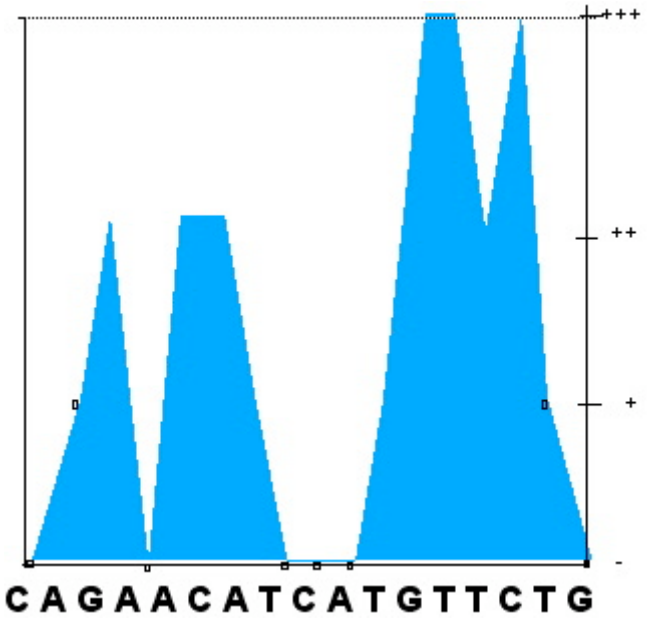

© 2004 by Genomatix Software GmbH

# Part I: Promoter elements

## General description of a transcription factor binding site

- A short stretch of DNA (about 10 to 25 nucleotides long)
- Bound by one (or more!) transcription factors
- TFBS are usually families of sequences, not defined strings
- TFBS contain highly conserved as well as variable nucleotides

Glucocorticoid receptor BS

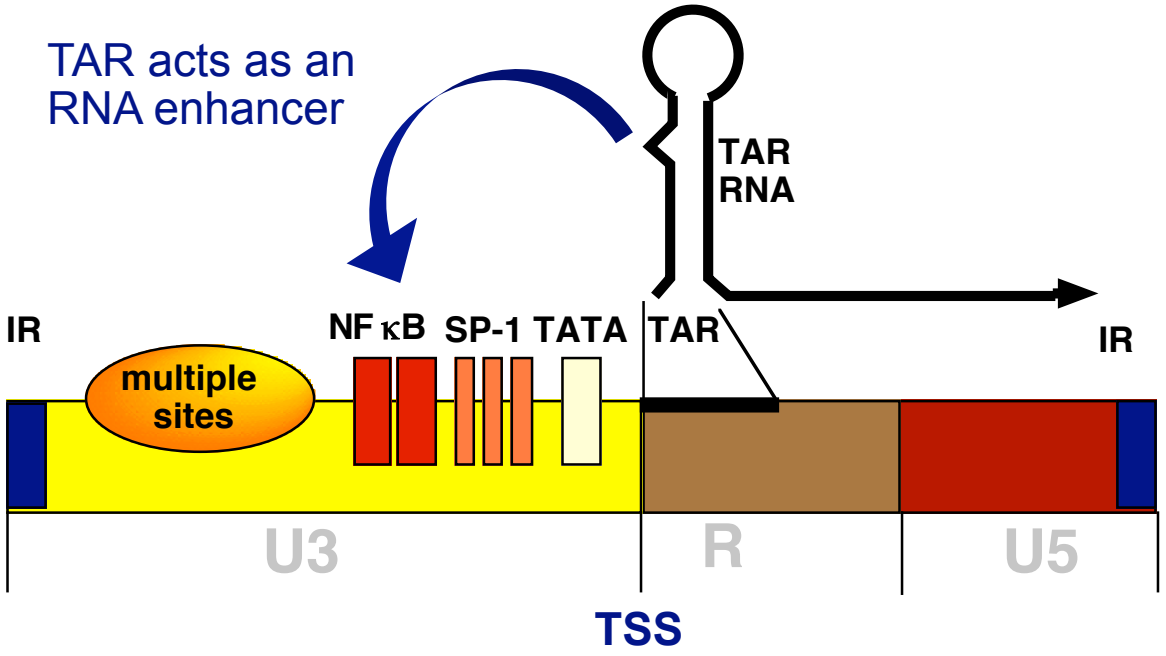



The inherent variability of TFBSs complicates specific recognition


© 2004 by Genomatix Software GmbH

# Part I: Promoter elements

## Polymerase II promoters may contain hairpin structures




The TAR hairpin is an essential structure in the HIV-1 promoter


© 2004 by Genomatix Software GmbH

# Part I: Promoter elements

## Summary

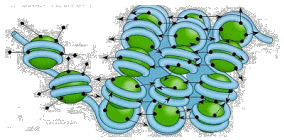
Promoter elements determine specificity and activity of promoters

- Transcription factor binding sites (TFBSs) are most important
- TFBSs can be detected in the primary DNA sequence
- TFBS are variable, not fixed sequences
- Repeats, hairpins are other known promoter elements

Promoters are composed of multiple elements

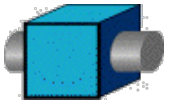

© 2004 by Genomatix Software GmbH

# Biological context of transcription



Chromatin organization

Transcription cycle



Promoter elements

**Understanding Promoter elements Part I**

Where to get promoter sequences?

Promoters in the genome

TFBS detection



© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## Distribution matrices allow determination of similarity

- Binding sites are aligned
- For each column the amount of each of the four nucleotides is counted
- Example: 6 sequences aligned, position 1 of alignment contains only A
- No arbitrary consensus
- Quantitative measure of similarity

| pos | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| A | 6 | 0 | 2 | 2 | 4 | |
| C | 0 | 3 | 2 | 2 | 0 | |
| G | 0 | 0 | 1 | 0 | 2 | |
| T | 0 | 3 | 1 | 2 | 0 | |
| | A | C | C | T | A | = 17 |
| | A | C | C | G | A | = 15 |
| | T | C | C | T | A | = 11 |

Distribution matrices do not have any intrinsic quality control!

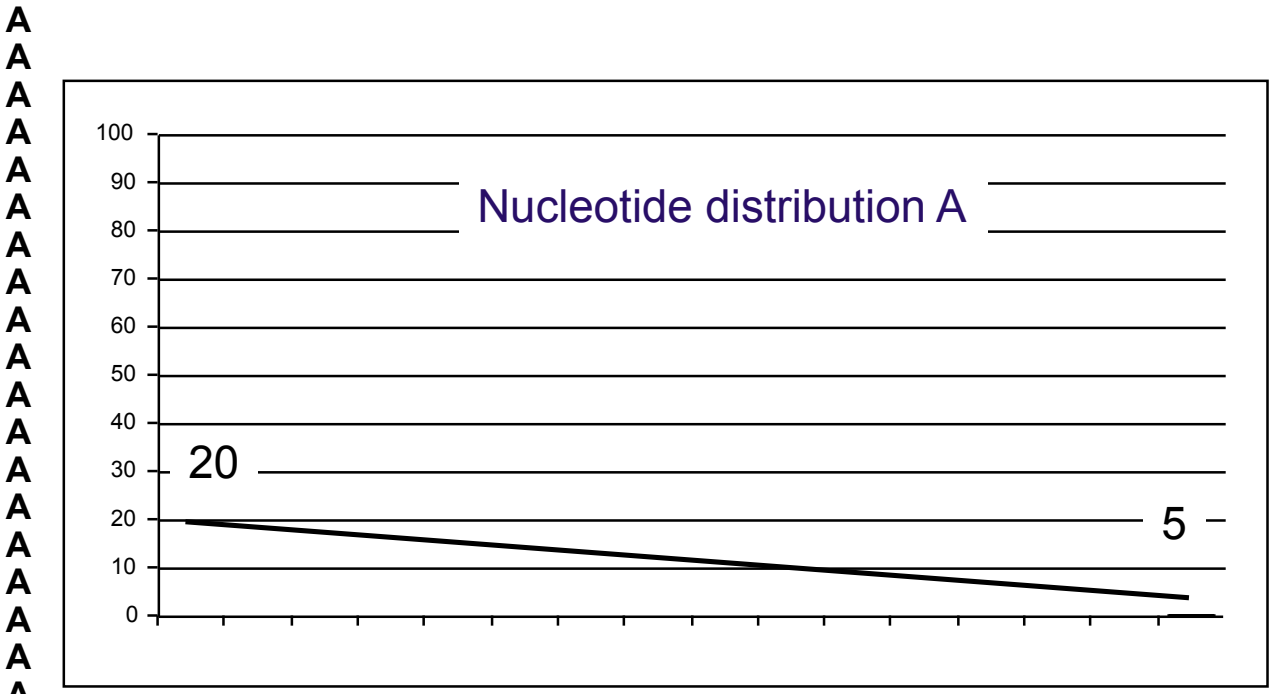


© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## Nucleotide distribution matrices

conserved




random

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

Maximum difference is 4-fold and depends on number of sequences!



© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## Nucleotide weight matrix using Shannon entropy

$$C = 100 / \ln 4 * \left( \sum_{\substack{b=A,C,G,T \\ 0,\ \text{if } p(b) = 0}} p(b) * \ln p(b) + \ln 4 \right)$$

| pos | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 6 | 0 | 2 | 2 | 4 |
| C | 0 | 3 | 2 | 2 | 0 |
| G | 0 | 0 | 1 | 0 | 2 |
| T | 0 | 3 | 1 | 2 | 0 |
| Ci | 100 | | | | |

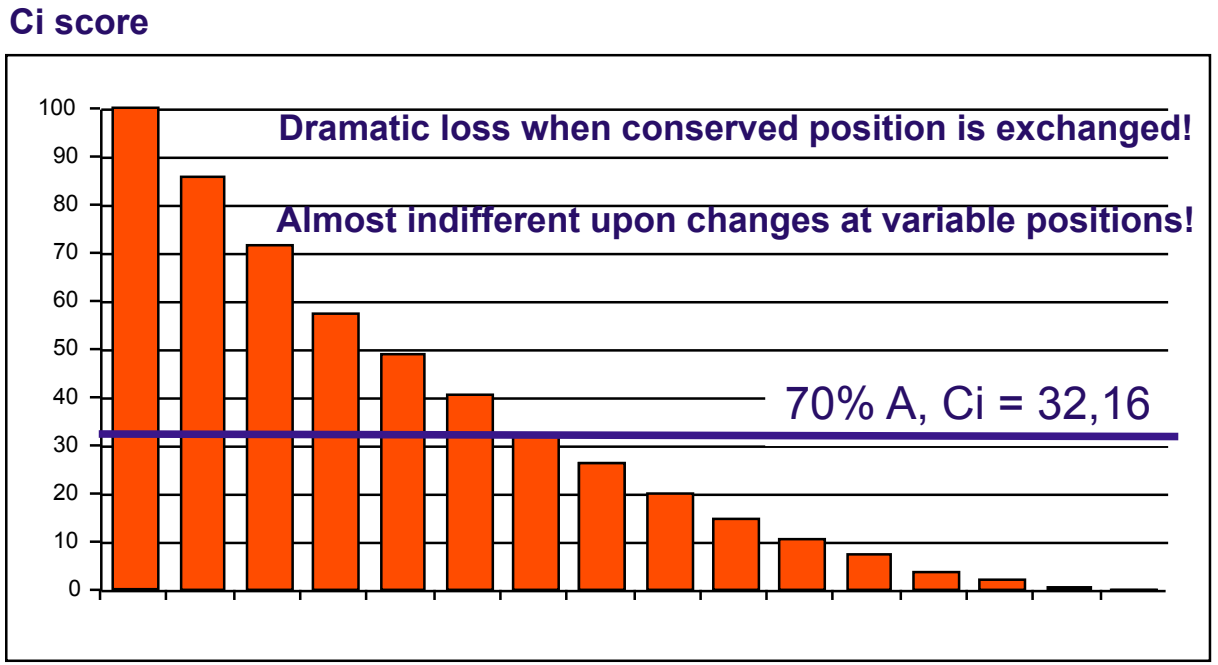
Inversion insures maximum value for conserved position!



© 2004 by Genomatix Software GmbH

## Effect of introducing Shannon entropy

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




Maximum scoring difference is now 100-fold!



© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## *MatInspector* scoring algorithm

| Matrix | pos. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | ◆14 | 0 | 0 | ◆15 | 0 | 2 | ◆9 | 3 | ◆11 | ◆8 | ◆22 | 0 | 0 |
| | C | 0 | 0 | ◆22 | 1 | ◆12 | 3 | 5 | 4 | 5 | 3 | 0 | ◆22 | 0 |
| | G | 8 | 0 | 0 | 4 | 0 | 4 | 3 | 3 | 3 | 5 | 0 | 0 | ◆22 |
| | T | 0 | ◆22 | 0 | 2 | 10 | ◆13 | 5 | ◆12 | 3 | 6 | 0 | 0 | 0 |
| | $C_i$: | **59** | **100** | **100** | **42** | **57** | **31** | **19** | **26** | **24** | **17** | **100** | **100** | **100** |
| Sequence | | **A** | **T** | **C** | **T** | **C** | **T** | **C** | **G** | **C** | **A** | **A** | **C** | **G** |

Numerator of mat_sim: $\Sigma$ (score * $C_i$) of the sequence □ = 13426

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| score | 14 * | 22 * | 22 * | 2 * | 12 * | 13 * | 5 * | 3 * | 5 * | 8 * | 22 * | 22 * | 22 * |
| $C_i$ | 59 | 100 | 100 | 42 | 57 | 31 | 19 | 26 | 24 | 17 | 100 | 100 | 100 |

Denominator of mat_sim: $\Sigma$ (maximum score * $C_i$) ◆ = 14426

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| score (maximum) | 14 * | 22 * | 22 * | 15 * | 12 * | 13 * | 9 * | 12 * | 11 * | 8 * | 22 * | 22 * | 22 * |
| $C_i$ | 59 | 100 | 100 | 42 | 57 | 31 | 19 | 26 | 24 | 17 | 100 | 100 | 100 |

**Matrix similarity = 13426/14426 = 0.93**



© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## *MatInspector* uses matrix as well as IUPAC libraries

| Group | # matrices | # matrix families | # IUPACs | # IUPAC families |
|---|---|---|---|---|
| Fungi | 42 | 31 | --- | --- |
| Insects | 37 | 24 | --- | --- |
| Plants | 67 | 39 | 323 | 258 |
| Vertebrates | 326 | 130 | --- | --- |
| Miscellaneous | 6 | 5 | --- | --- |
| Other Functional Elements | 7 | 4 | --- | --- |
| **Total** | **485** | **233** | **323** | **258** |

The **vertebrate** group represents binding site descriptions of **1000 TFs covered**!

- Homo sapiens: 406
- Mus musculus: 373
- Rattus norvegicus: 168
- Danio rerio: 53



© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## Why matrix families?

- There are several matrices for the same or very similar factors
  (e.g. *cAMP responsive elements*: V$CREB, V$ATF, V$P300, V$JUN)
- Programs to locate binding sites report several matches
  at the same or very similar positions
- Result lists are inflated and are hard to interpret
- There is no such thing as the “best” matrix to be used exclusively

More binding sites reported does not mean better quality of prediction

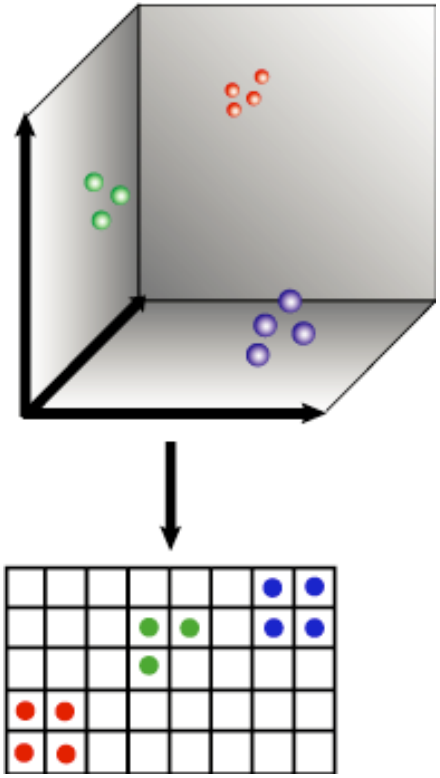

© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## Self Organizing maps (SOMs)

SOMs are also known as Kohonen cards

- Data described by many different features can be arranged in two dimensions
- Projection of feature vectors
- This way clusters of similar data are easily detected

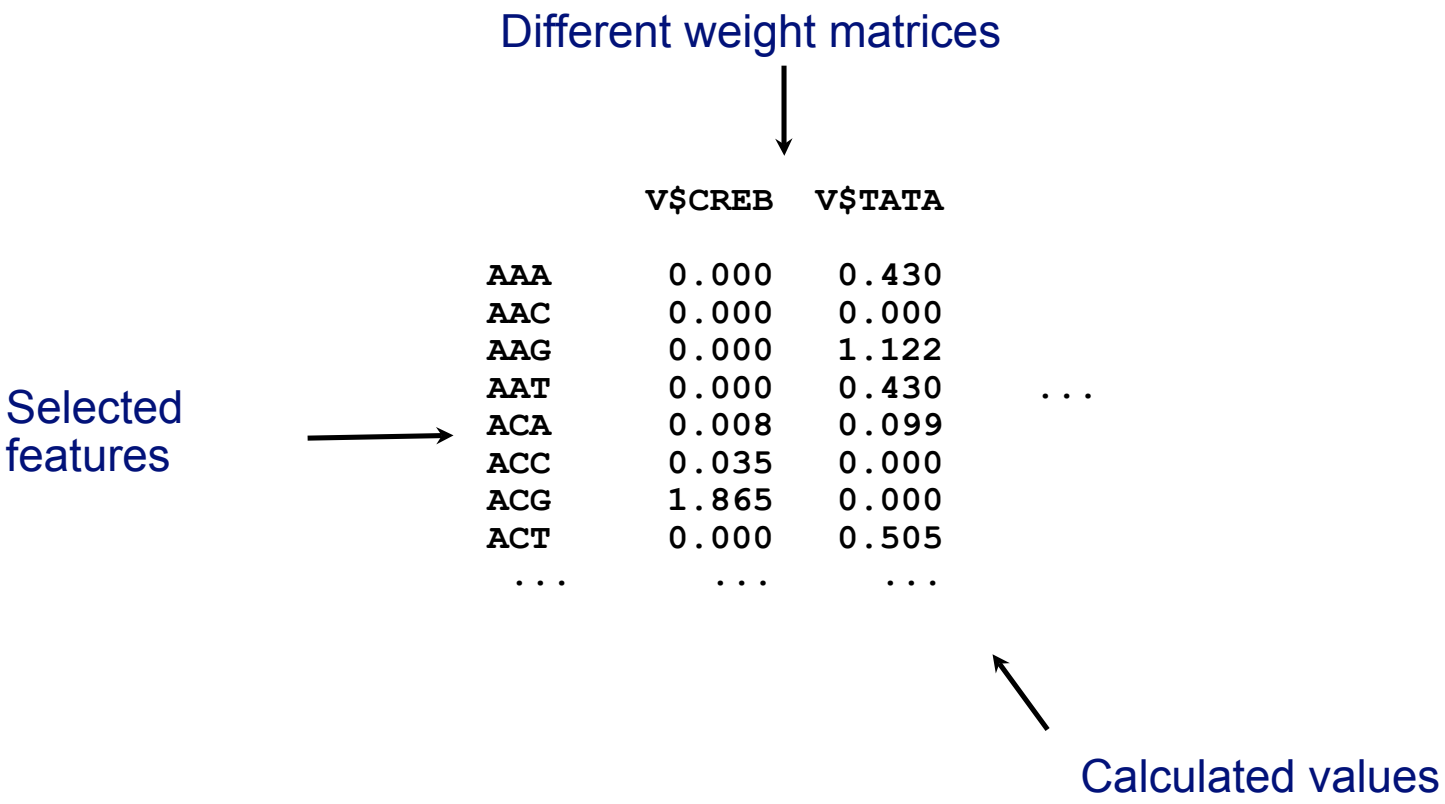




© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## A statistical approach...

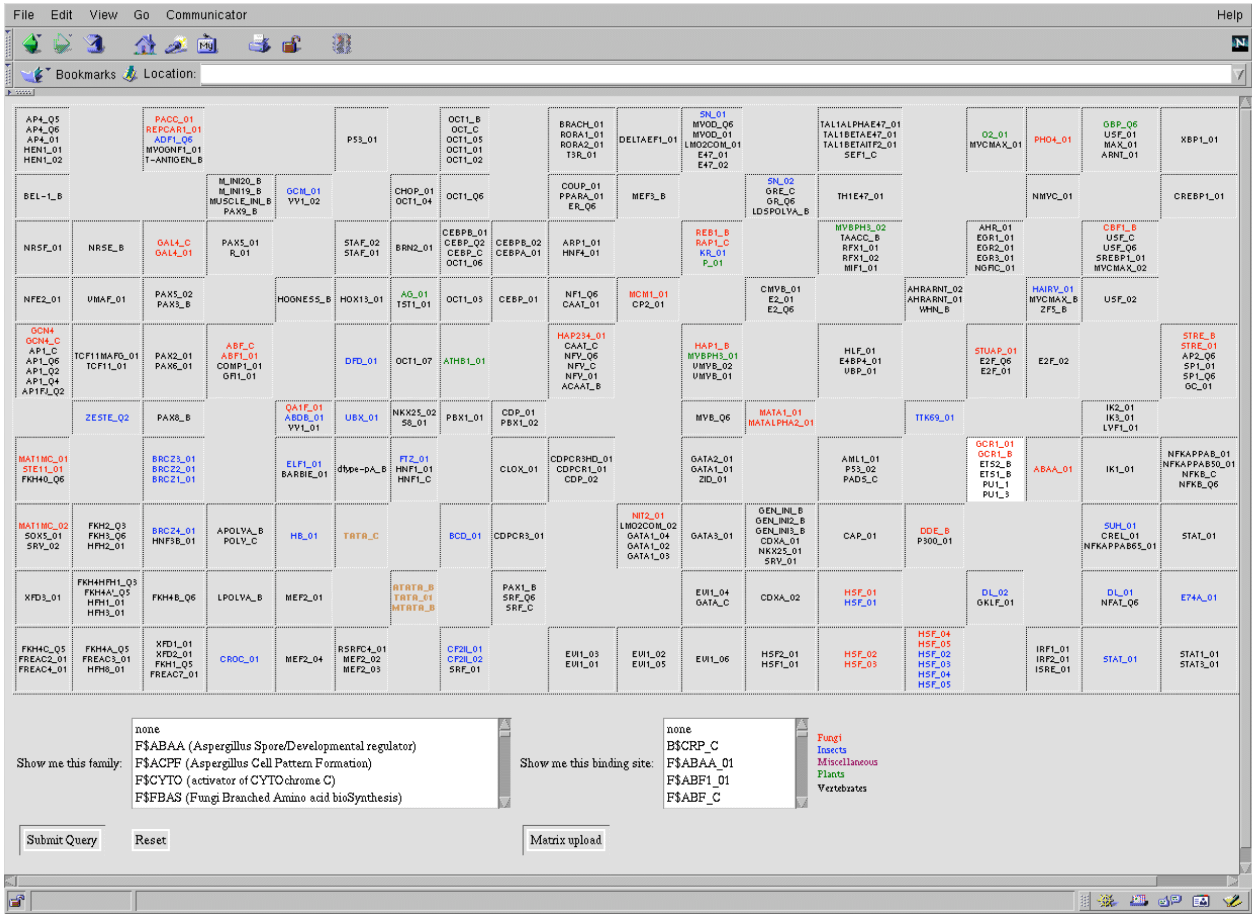




© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## ...turns out to be a useful concept!

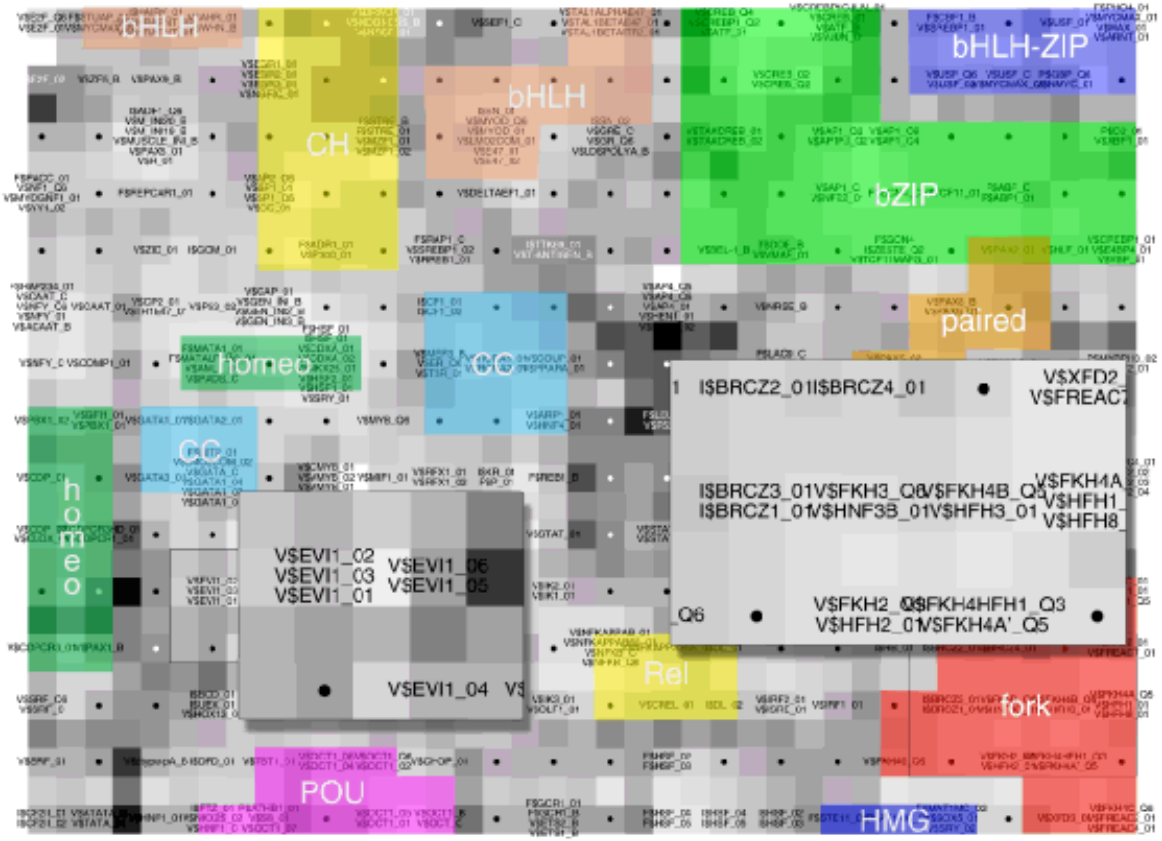





© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## *MatInspector* matrix families agree with protein data




There is obviously a general protein-DNA recognition code!



© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

- Analysis with individual matrices

**Inspecting sequence hivth475a [L31963] (1 - 672):**

| Name of family/matrix | Position | Strand | Core sim. | Matrix sim. | Sequence |
|---|---|---|---|---|---|
| V$NFKB_Q6 | 345 | (+) | 1.000 | 0.968 | aaGGGActttccgc |
| V$NFKB_C | 346 | (+) | 1.000 | 0.977 | aGGGACtttccg |
| V$NFKAPPAB_01 | 347 | (+) | 1.000 | 0.988 | GGGActttcc |
| V$NFKAPPAB65_01 | 347 | (+) | 1.000 | 0.984 | gggactTTCC |
| [illegible] | [illegible] | (+) | 1.000 | 0.994 | tgGGGActttccag |
| V$NFKB_C | 360 | (+) | 1.000 | 1.000 | gGGGACtttcca |
| V$NFKAPPAB_01 | 361 | (+) | 1.000 | 0.988 | GGGActttcc |
| V$NFKAPPAB65_01 | 361 | (+) | 1.000 | 0.984 | gggactTTCC |
| V$NFKB_Q6 | 400 | (+) | 1.000 | 0.994 | tgGGGActttccag |
| V$NFKB_C | 401 | (+) | 1.000 | 1.000 | gGGGACtttcca |
| V$NFKAPPAB_01 | 402 | (+) | 1.000 | 0.988 | GGGActttcc |
| V$NFKAPPAB65_01 | 402 | (+) | 1.000 | 0.984 | gggactTTCC |

- Analysis with matrix families

**Inspecting sequence hivth475a [L31963] (1 - 672):**

| Name of family/matrix | Position | Strand | Core sim. | Matrix sim. | Sequence |
|---|---|---|---|---|---|
| V$NFKB/NFKAPPAB_01 | 347 | (+) | 1.000 | 0.988 | GGGActttcc |
| V$NFKB/NFKB_C | 360 | (+) | 1.000 | 1.000 | gGGGACtttcca |
| V$NFKB/CREL_01 | 388 | (+) | 1.000 | 0.865 | ggaactTTCC |
| V$NFKB/NFKB_C | 401 | (+) | 1.000 | 1.000 | gGGGACtttcca |


© 2004 by Genomatix Software GmbH

# Part I: TFBS detection

## Summary

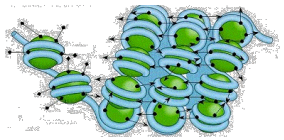
TFBS can be described and detected by mathematical models

- Matrix-based approaches are more specific than IUPAC strings
- Nucleotide distribution matrices have a small scoring range
- Weight matrices can successfully predict binding affinity
- Matrix families account for the biological variability of TFs

TFBSs can be reliably predicted by *in silico* approaches

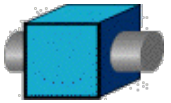

© 2004 by Genomatix Software GmbH

# Biological context of transcription



Chromatin organization

Transcription cycle



Promoter elements

**Understanding Promoter elements Part I**

Where to get promoter sequences?

Promoters in the genome

| pos | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 6 | 0 | 2 | 2 | 4 |
| C | 0 | 3 | 2 | 2 | 0 |
| G | 0 | 0 | 1 | 0 | 2 |
| T | 0 | 3 | 1 | 2 | 0 |

TFBS detection

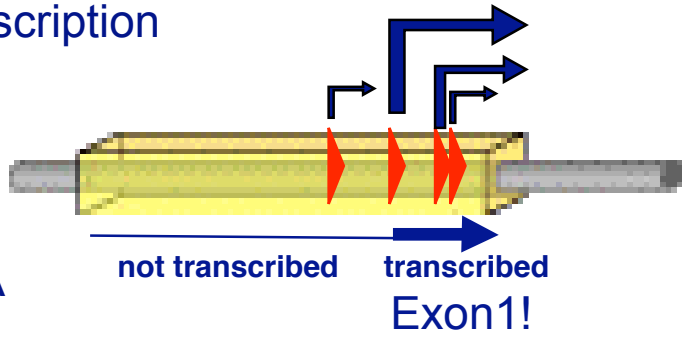


© 2004 by Genomatix Software GmbH

# Part I: Promoters in the genome

## General description of a polymerase II promoter region

- A short stretch of DNA containing a transcription start site (TSS)
- Capable to support initiation of transcription
- Most of a polymerase II promoter region is not transcribed
- Promoter regions are genomic DNA
- Promoter regions may contain multiple TSS
- Promoter regions are not necessarily connected to a coding sequence
- Promoter regions vary in length from 150 bp to 2,500 bp

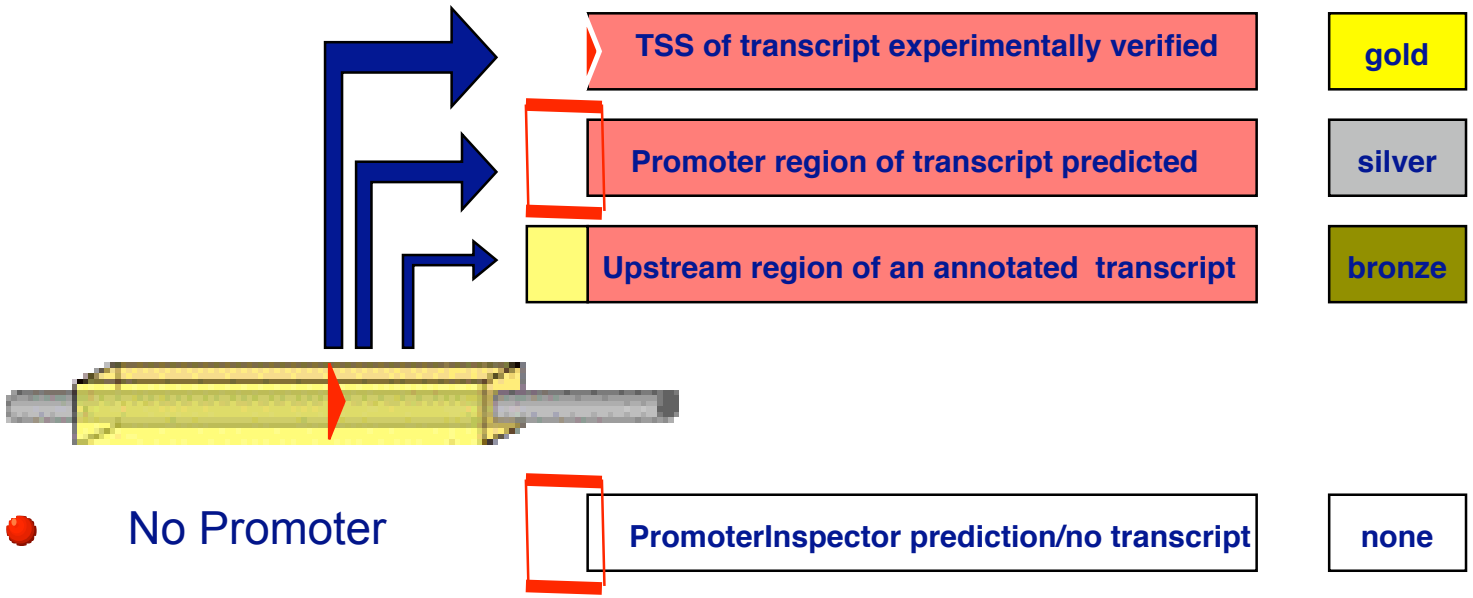



The first exon of every mRNA is or contains part of a promoter sequence!


© 2004 by Genomatix Software GmbH

# Part I: Promoters in the genome

## Definition of promoter regions and transcript quality



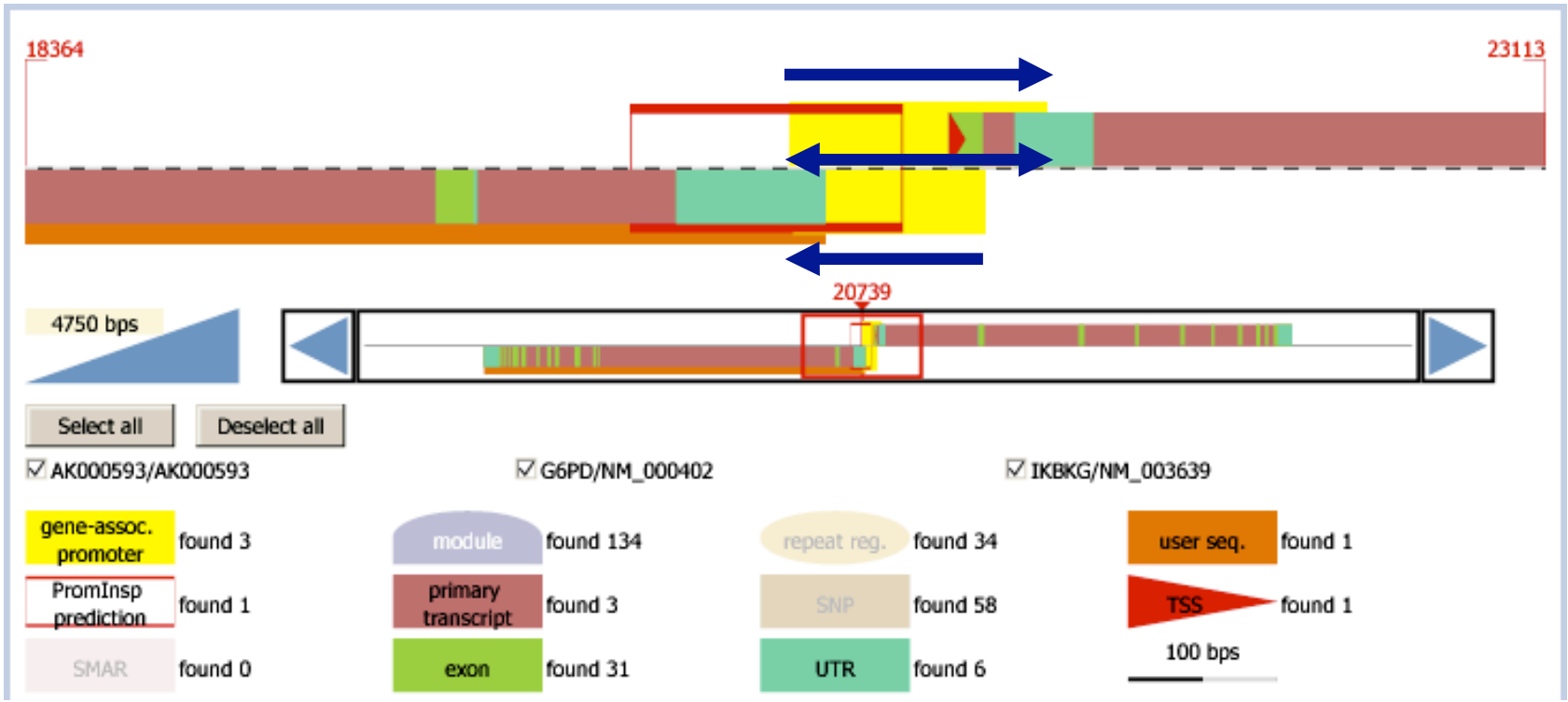

- No Promoter
- Each transcript is associated with a quality
- No quality can be assigned to the promoter region

A promoter region solely defines a sequence range in the genome


© 2004 by Genomatix Software GmbH

# Part I: Promoters in the genome

## Promoter regions can be multifunctional




Is this a bidirectional promoter or are these two independent promoters?



© 2004 by Genomatix Software GmbH

# Part I: Promoters in the genome

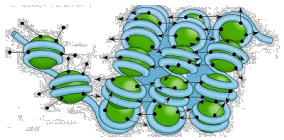
## Pol II promoters defy “simple” rules of sequence analysis

- ~~Functional similarity is reflected in sequence similarity~~
  Promoters as well as their elements show low sequence conservation
- ~~Promoters have a general common structure~~
  Promoters need to be different to ensure different regulation
- ~~Promoters are simple clusters of promoter elements~~
  Promoters require a defined internal organization to be functional
- ~~Locating transcription factor binding sites indicates promoters~~
  All promoter elements can be found almost everywhere in DNA

Only the complete context of elements determines a promoter function!

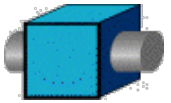

© 2004 by Genomatix Software GmbH

# Biological context of transcription



Chromatin organization

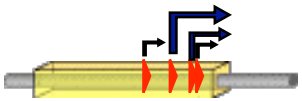
Transcription cycle



Promoter elements

**Understanding Promoter elements Part I**

Where to get promoter sequences?



Promoters in the genome

| pos | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 6 | 0 | 2 | 2 | 4 |
| C | 0 | 3 | 2 | 2 | 0 |
| G | 0 | 0 | 1 | 0 | 2 |
| T | 0 | 3 | 1 | 2 | 0 |

TFBS detection



© 2004 by Genomatix Software GmbH

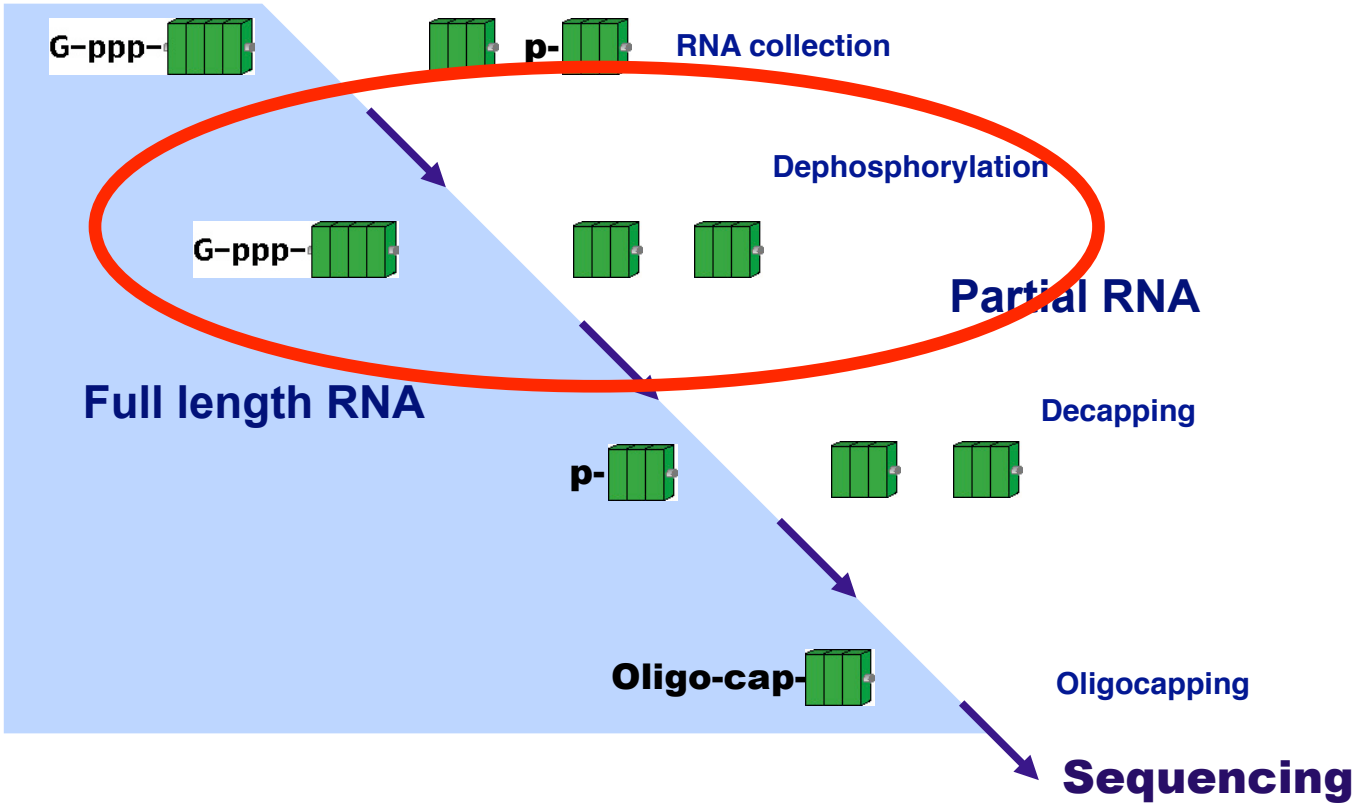
## How to get to mammalian promoter sequences?

- Cloning of sequence, verification in transient expression assays — **Most reliable**
- Exact mapping of 5' end of *complete* mRNA to genomic DNA — **Reliable**
- Phylogenetic conservation of known promoter from other species — **Reliable**
- Prediction of promoter by *suitable* bioinformatics methods — **Reliable**
- Phylogenetic conservation of transcript from other species — **Acceptable**
- Upstream region of an annotated transcript — **Dangerous**
- Upstream region of a cDNA (usually 5' incomplete!) — **Very dangerous**

Mammalian promoter sequence determination is still a difficult task!


© 2004 by Genomatix Software GmbH

## Identification of true 5'-ends of mRNAs by oligo-capping




Dephosphorylation is only 70% to 80% complete!


© 2004 by Genomatix Software GmbH

# Genomatix Part I: Where to get promoter sequences

## Genomatix annotated promoters in the human, mouse, and rat genomes

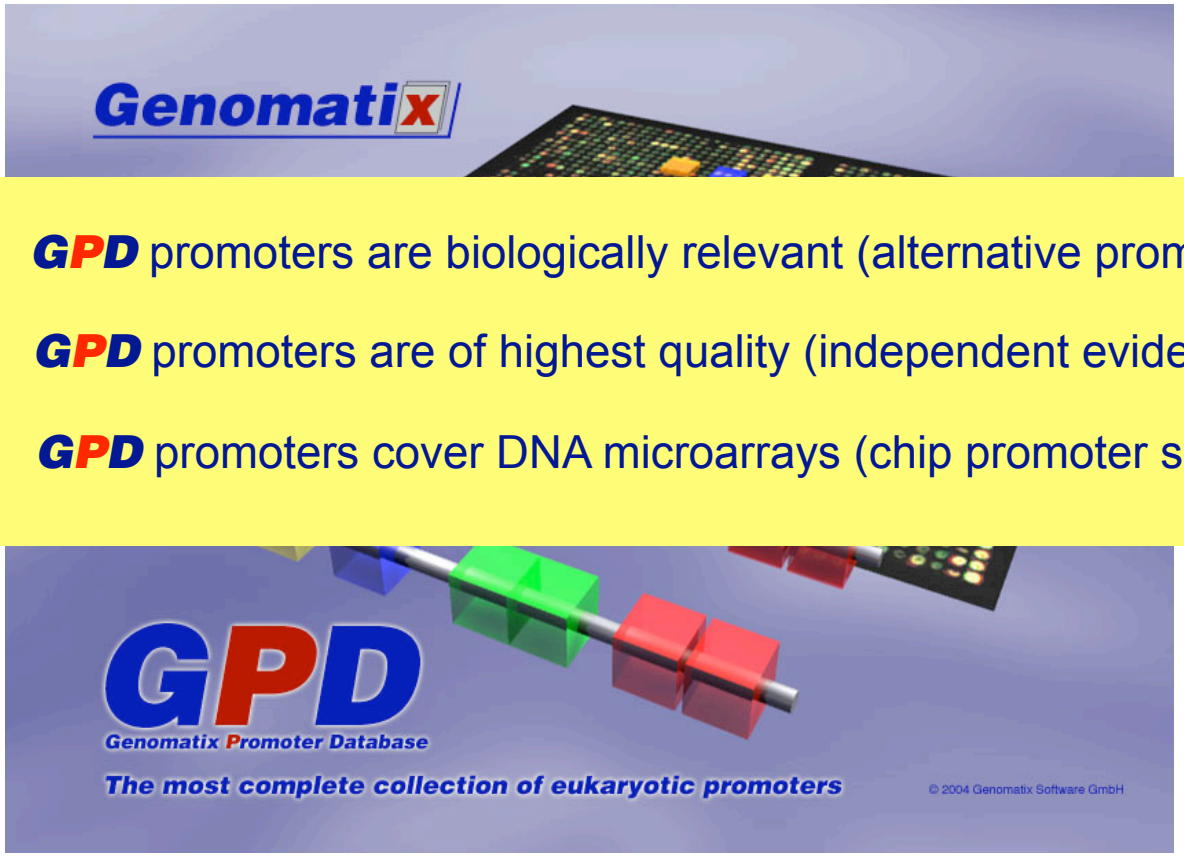
The Genomatix Genome Engine (***ElDorado***)

- 50,145 gene-associated human promoters (28,538 exp. verified)
- 77,424 gene-associated mouse promoters (55,160 exp. verified)
- 33,392 gene-associated rat promoters (3,319 verified by prediction)
- Length of promoter regions: ~ 600 bp (user-defined)
- ~210 GB total sequence & annotation compiled
- Open for free registration (monthly limits)

***ElDorado*** currently contains ~30,000 human annotated genes with promoters


© 2004 by Genomatix Software GmbH

## GPD provides complete promoter sets for several genomes




- **GPD** promoters are biologically relevant (alternative promoters)
- **GPD** promoters are of highest quality (independent evidence)
- **GPD** promoters cover DNA microarrays (chip promoter sets)

www.promoter-resource.com


© 2004 by Genomatix Software GmbH

# Part I: Promoters in the genome

## Summary

Promoters are defined by their transcriptional function

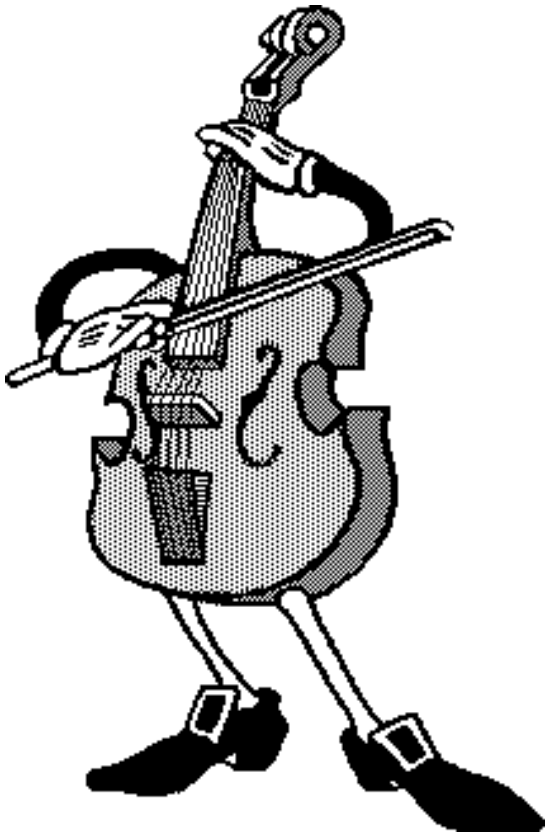
- Genomic sequences contain promoter regions
- A single promoter region can initiate several transcripts
- Promoter analysis requires other approaches than protein analysis
- Promoter regions are available for several complete genomes

A promoter region is just a piece of DNA sequence


© 2004 by Genomatix Software GmbH

# End of part one

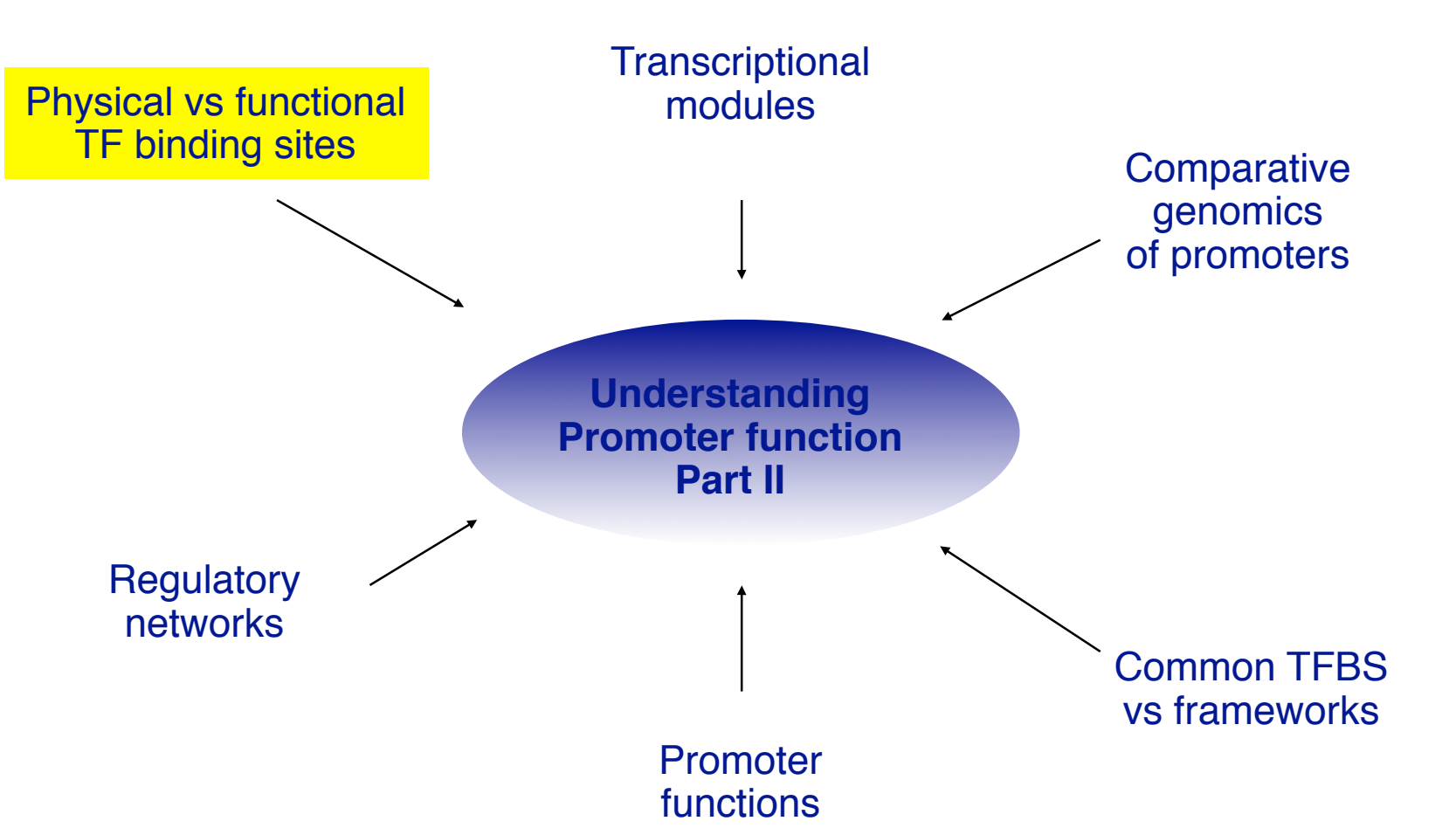
## Time for a break!





© 2004 by Genomatix Software GmbH

# Genomatix Functional organization of transcription





© 2004 by Genomatix Software GmbH

# Part II: Physical vs functional TFBS

## Transcription factor binding sites (TFBS) are ubiquitous

TFs do not discriminate between promoters and other regions in mammals

- TFBS Max: ≈ 7,700 matches in human genome
- TFBS SP1: ≈ 415,000 matches in human genome
- TFBS TATA-box: ≈ 5,200,000 matches in human genome
- There are less than 40,000 genes and only half have a TATA-box
- Even piles of TF binding sites are not indicative of a promoter

Mammalian TF binding sites become functional only as organized groups


© 2004 by Genomatix Software GmbH

# Part II: Physical vs functional TFBS

## TFBSs are not generally overrepresented in human promoters

TFs do not discriminate between promoters and other regions in mammals

- 127 ***MatInspector*** matrix families represent more than 1000 TFs
- Statistically, TFBS families are 1.4 fold overrepresented in human promoters as compared to the whole human genome
- Only 28 out of 127 TF families reach or exceed that limit (22%)
- 73 are actually underrepresented in human promoters (57%)

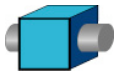
Frequency of TFBSs depends to a large extent on local GC content


© 2004 by Genomatix Software GmbH

# Part II: Physical vs functional TFBS

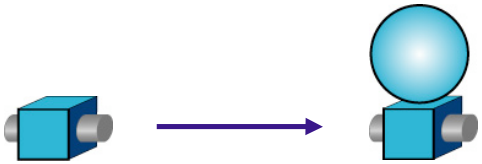
## A physical binding site is invariable

- A physical binding site is a fixed part of the genome



= weight matrix / IUPAC string

- Physical binding sites can be detected by ***MatInspector***
- This DNA sequence usually can bind to its cognate protein(s)



Physical binding sites have no function in transcription on their own



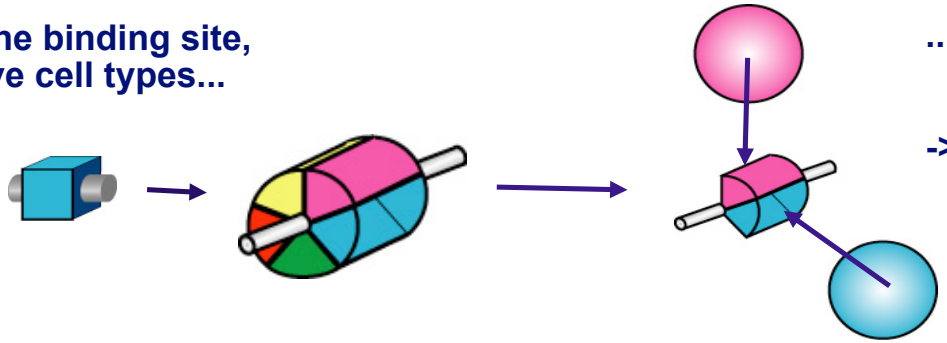
© 2004 by Genomatix Software GmbH

# Part II: Physical vs functional TFBS

## A functional binding site depends on context!

- A functional binding site requires a cellular context

**One binding site, five cell types...**

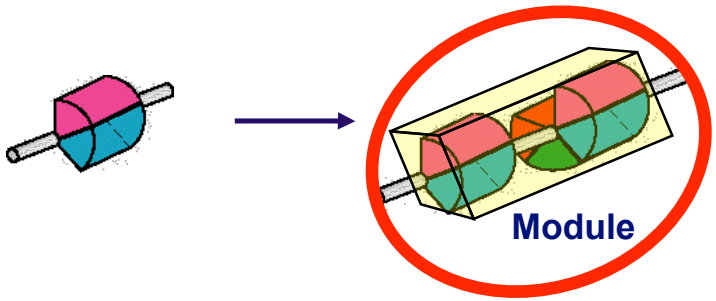


**...but binding proteins are present only in 2 cell types!**

**-> no *functional* binding site in the other 3 cell types!**

- A functional binding site requires a genomic context

**Even when binding proteins are present...**




**...biological function may require additional binding sites!**

Transcriptional function is defined by the cellular and genomic context



© 2004 by Genomatix Software GmbH

# Genomatix Functional organization of transcription

Physical vs functional
TF binding sites

Transcriptional modules

Comparative genomics of promoters

**Understanding Promoter function Part II**

Regulatory networks

Promoter functions

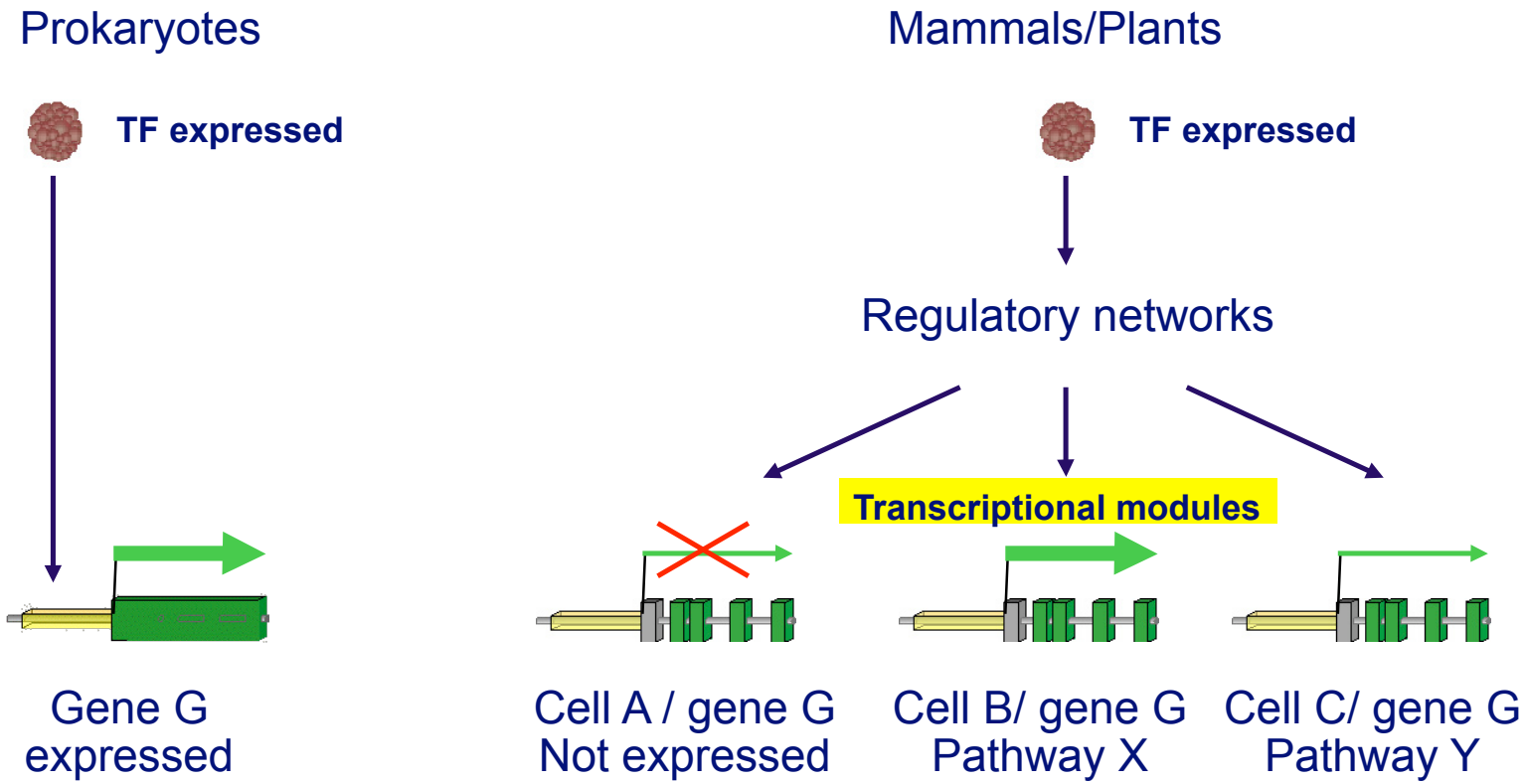
Common TFBS vs frameworks



© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

## Single cells versus differentiated multicellular organisms




Response to TFs is cell-specifically differentiated in multicellular organisms

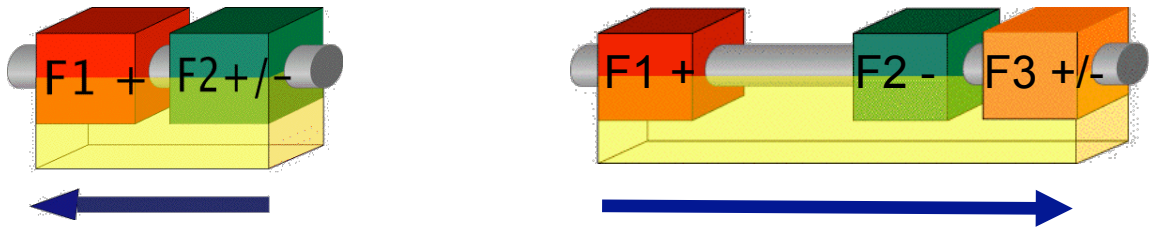


© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

## A transcriptional module is the smallest functional unit

- A transcriptional module consists of two or more TFBSs
- Strand orientation, relative order and distance of TFBSs are important
- A module also has a strand orientation and can shift within a promoter
- Transcriptional modules are present in promoters and enhancers




Transcriptional modules integrate signals via the interacting Transcription Factors


© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

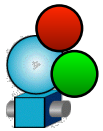
## Transcription factor (TF) binding modes

- Direct binding (affinity driven)

**MatInspector**

$S_{binding} = f(TFBS_{affinity})$

- Indirect binding (protein-mediated)



$S_{binding} = f_1(TFBS_{affinity} * f_2(S_{n(interacting\ factors)}))$

- Cooperative binding (protein-mediated by another TF)

$S_{binding(A)} = f_1(TFBS_{affinity(A)} * f_2(TFBS_{affinity(m)} * f_3(S_{m(interacting\ factor\ m)}))$

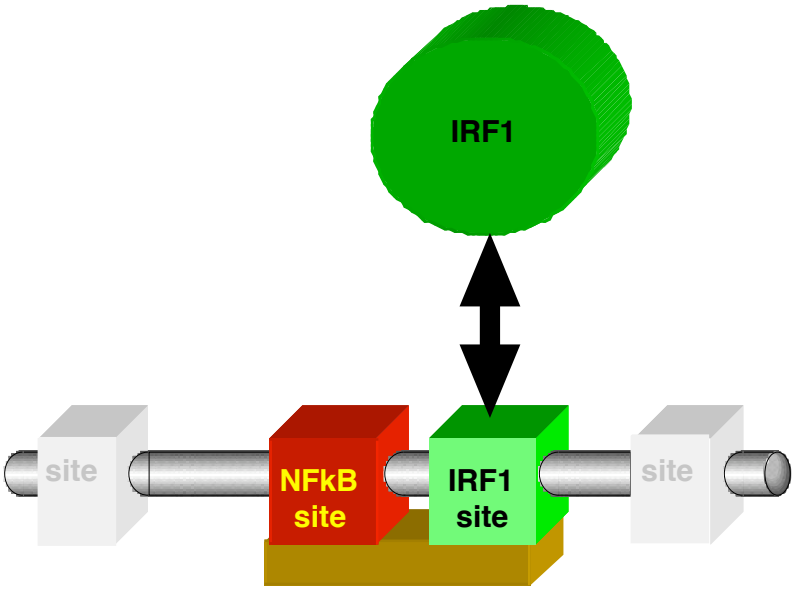
Protein-mediated binding depends more on context than binding site affinity!



© 2004 by Genomatix Software GmbH

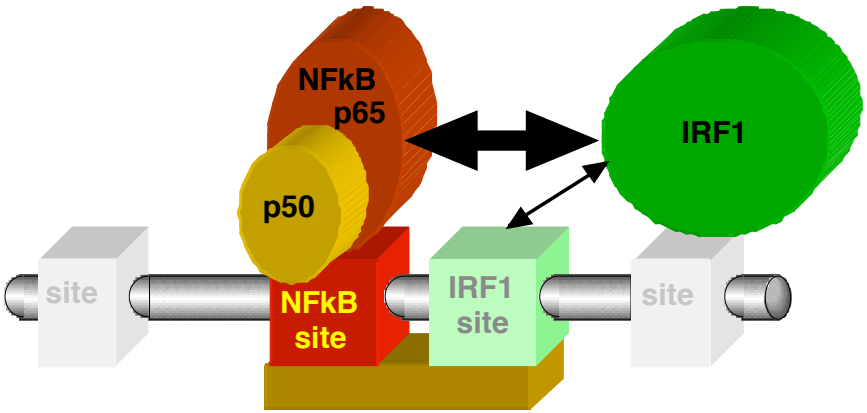
# Part II: Transcriptional modules

## The context of a promoter module can allow for one weak site




High score / High score

Direct DNA binding to strong site




High score / Low score

Cooperative binding with other protein

Promoter modules cannot be found by simple combination of sites



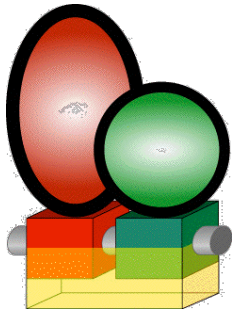
© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

## Promoter modules can work in three different ways

### Synergistic

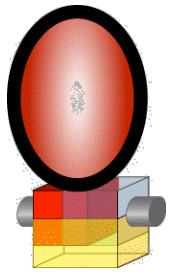
**"short range module"**
**distance ≤ 50 bp**



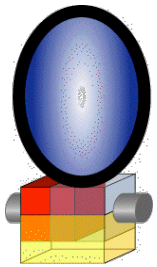
Binding Affinity: **High / Low Is possible**

### Antagonistic

**"Composite elements"**

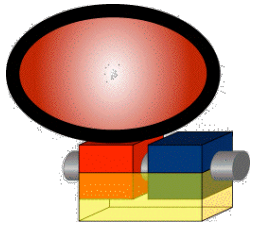


or

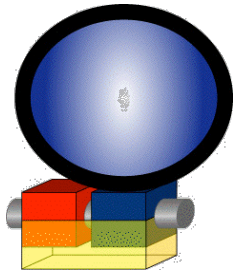


Binding Affinity: **High / Low Is possible**

**"short range module"**
**distance ≤ 50 bp**



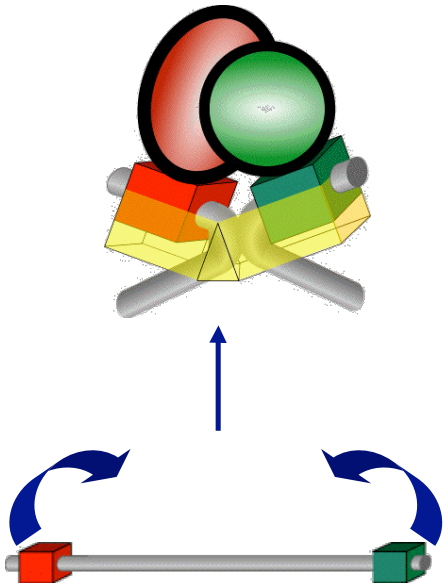
or



Binding Affinity: **High / Low Is possible**

### Synergistic

**"Looping module"**
**distance up to 300bp**



Binding Affinity: **High / High only**

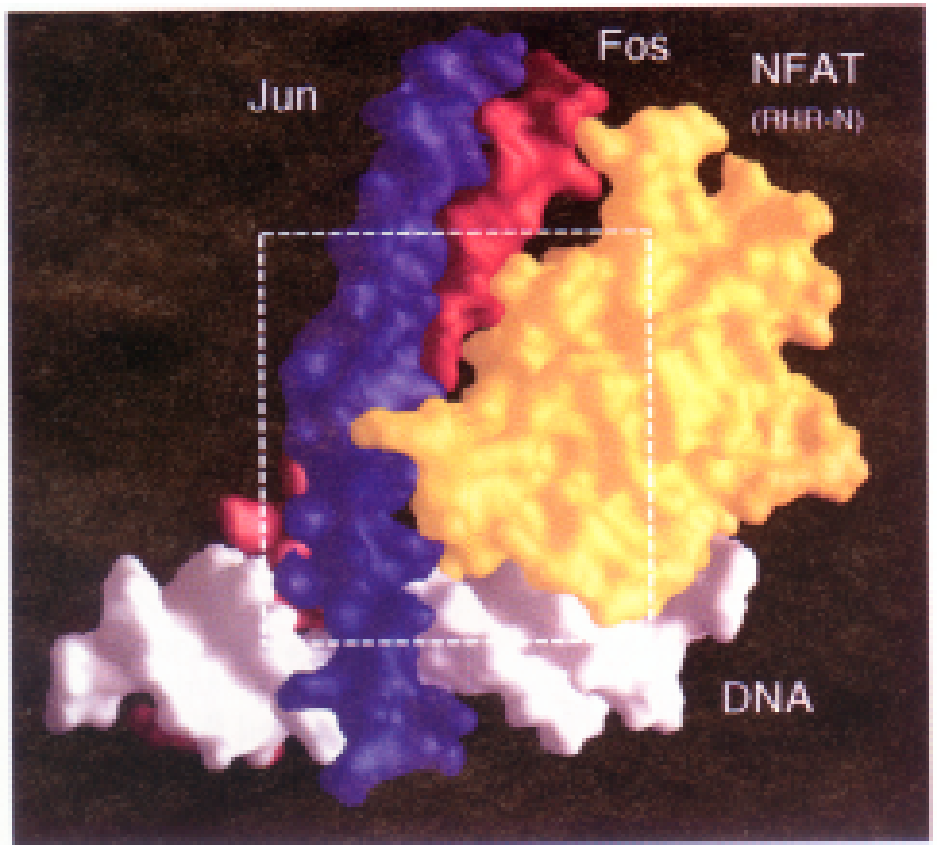


© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

## The DNA-protein complex of the NFAT-AP1 promoter module

- The **strand orientation** of elements must support 3-D protein complexes
- The **relative order** of elements can be crucial
- The **distance** of elements is important



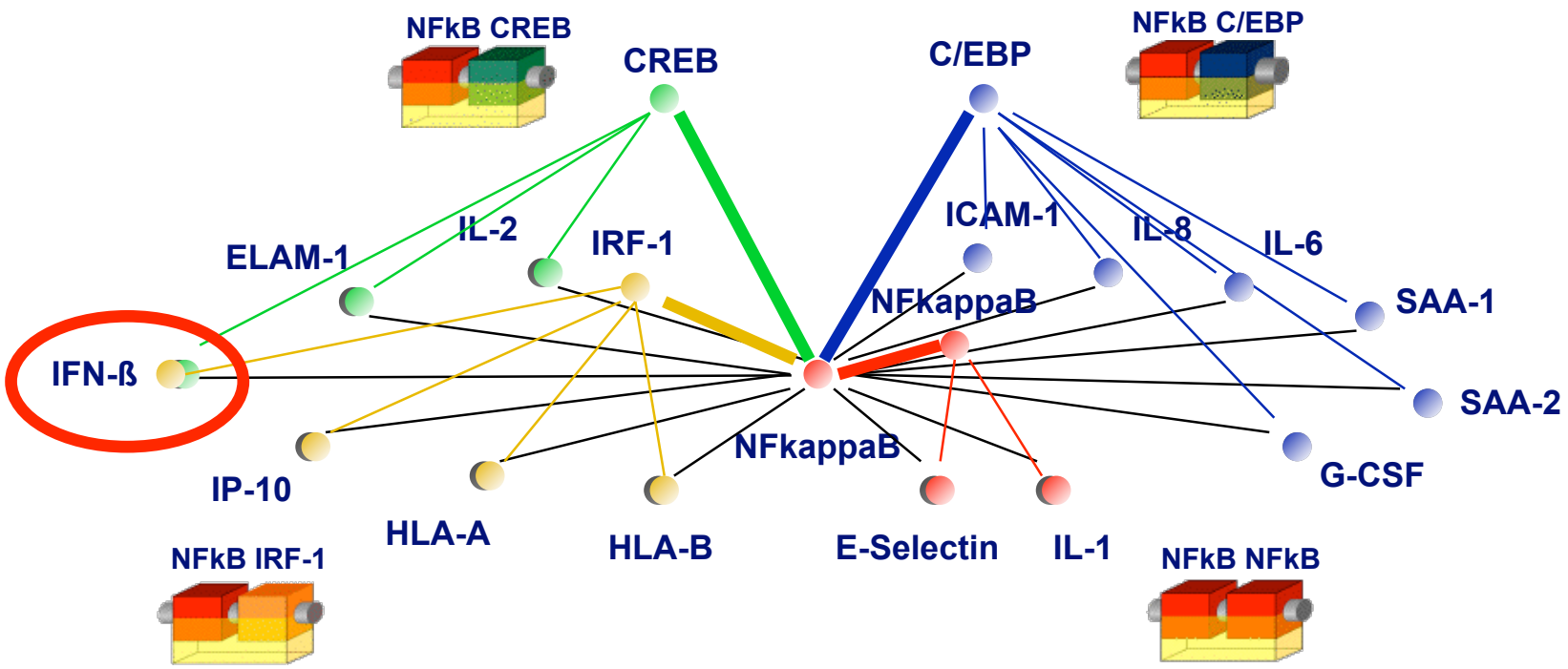


© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

## Transcriptional modules define target genes of pathways

- NFkappaB is involved in regulation of target genes of several pathways




Modules are the basic elements of regulatory pathways and networks

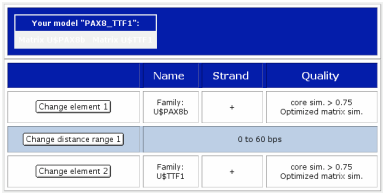


© 2004 by Genomatix Software GmbH

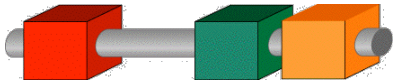
# Part II: Transcriptional modules

## Describing promoter/enhancer organization

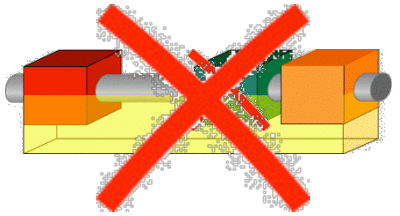
- A ***model*** is a mathematical description of any kind of promoter / enhancer organization




- A ***framework*** describes a conserved set of transcription elements



- A ***module*** describes experimentally verified mutually dependent functional transcription elements



**Delete a single site - lose function of all**

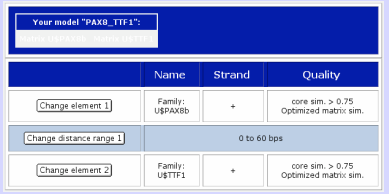
Modules definitely have a biological function, frameworks may have a function


© 2004 by Genomatix Software GmbH

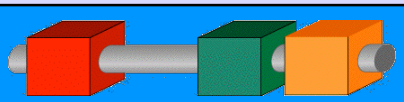
# Part II: Transcriptional modules

## Describing promoter/enhancer organization

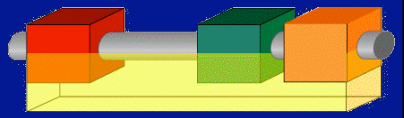
*model*




*framework*



*module*



(requires experimental verification)

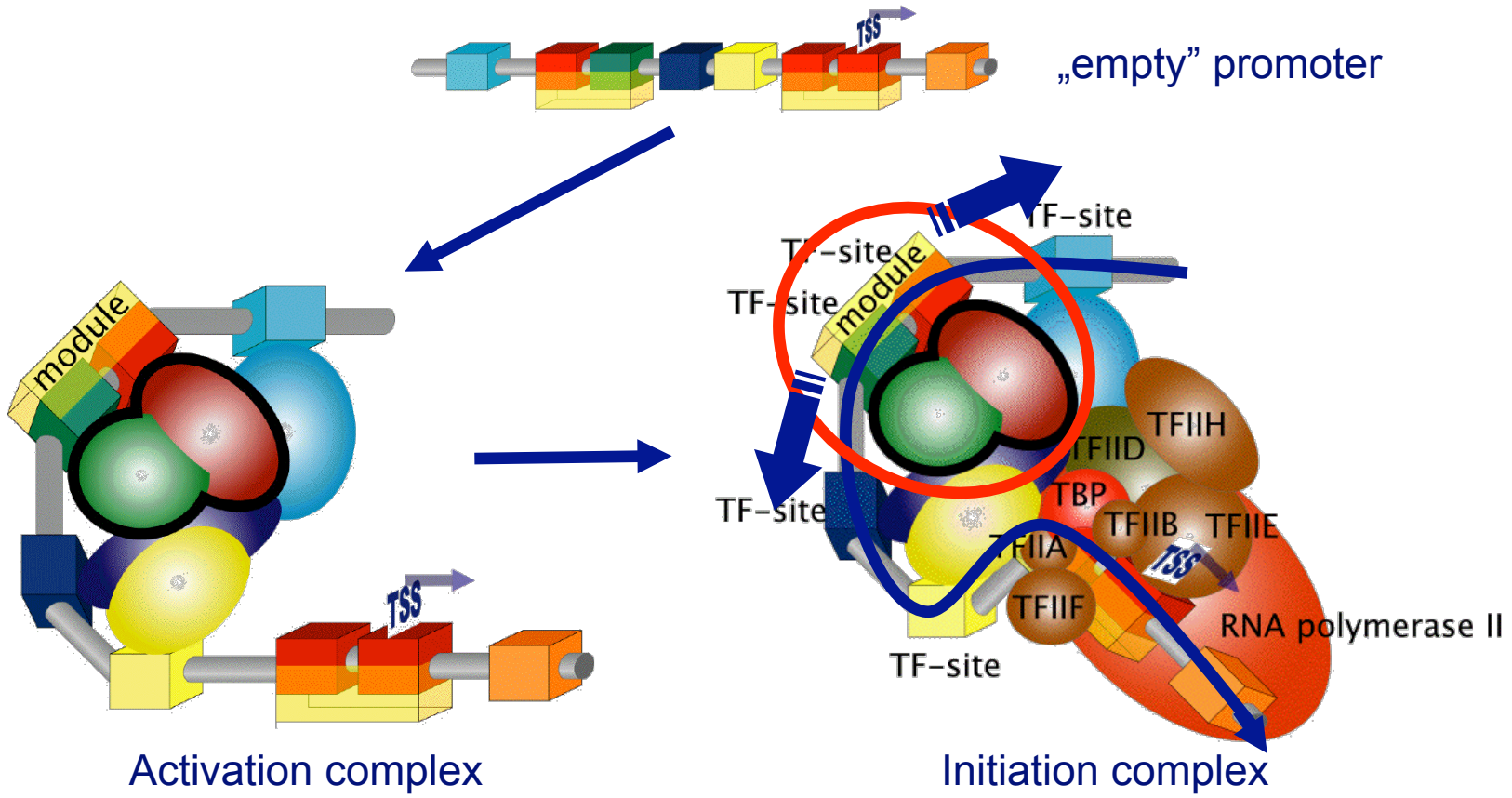
Models, frameworks and modules are hierarchically connected



© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

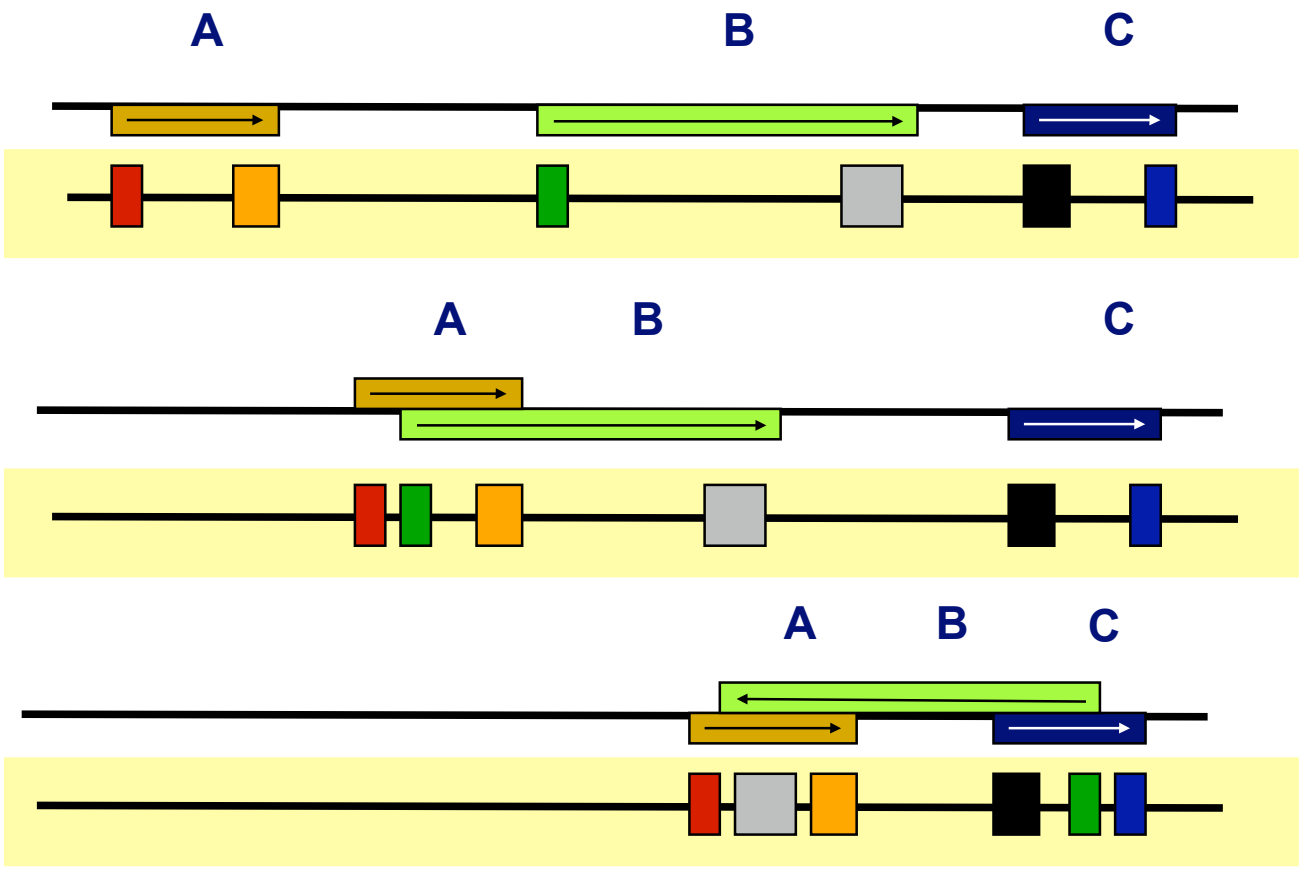
## Promoter modules guide complex formation on promoters






© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

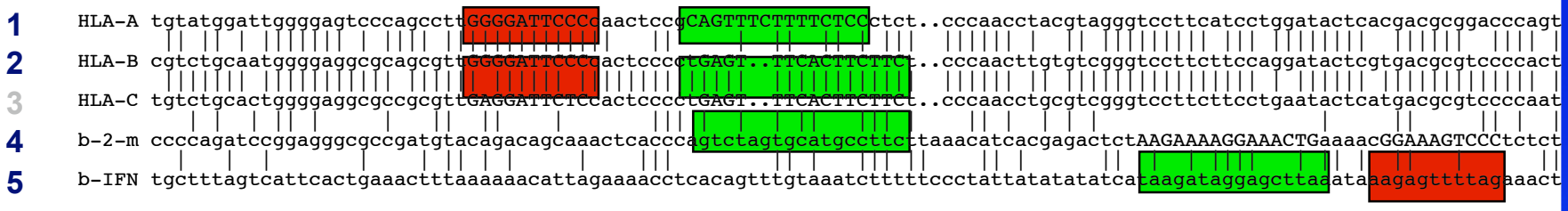


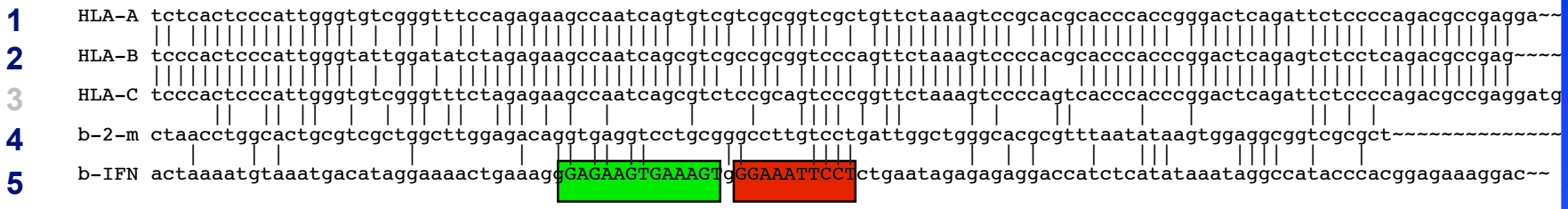

No common organization? → Common modules!


© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

## All 5 promoters, except for #3, are induced by TNF and IFN




The module consists of two transcription factor binding sites (NFκB and IRF)

BIOINFORMATICS

Klingenhoff et al., (1999)Bioinformatics 15, 180-186

The promoter module shifts and changes strand orientation in these promoters



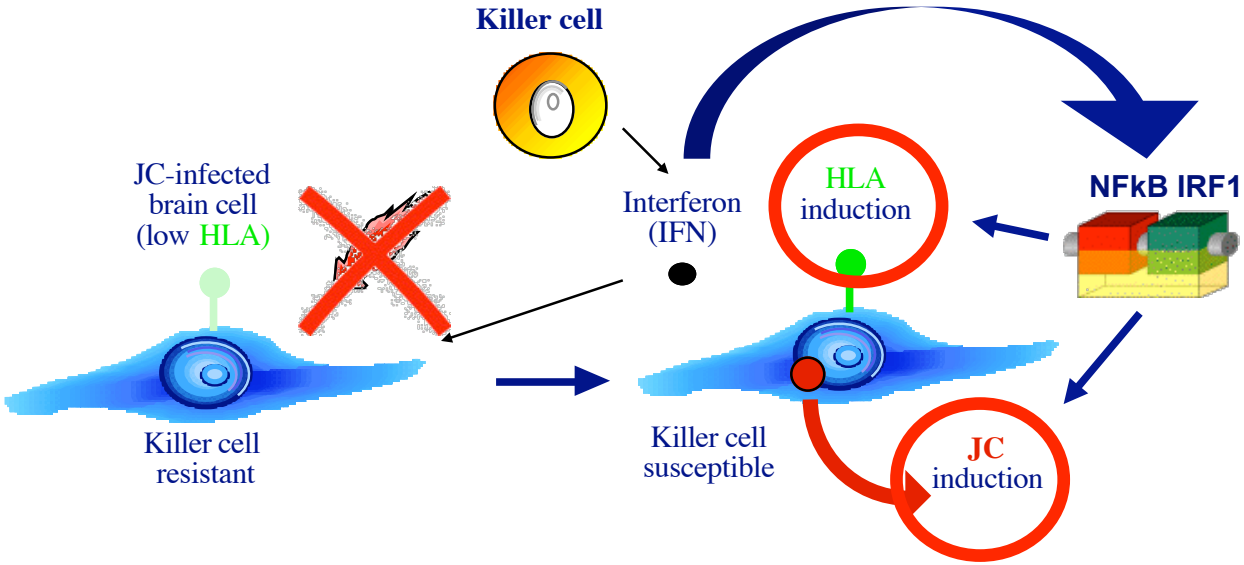
© 2004 by Genomatix Software GmbH

# Part II: Transcriptional modules

## The TNF/IFN module reveals molecular basis for viral escape!

**JC virus infects brain cells with low HLA class I expression, causing progressive multifocal leukoencephalopathy (PML)**

**(Problem in immunosupressed patients like AIDS patients)**




…and JC virus probably uses the same signal to escape!


© 2004 by Genomatix Software GmbH

## Summary

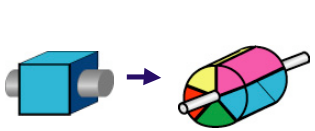
TFBSs can assume several functions depending on the context

- Individual TFBS only bind to TFs, but have no transcriptional function
- The smallest functional unit in transcription is a transcriptional module
- Transcriptional modules exhibit a strict internal organization
- Transcriptional modules can move and flip strand in promoters

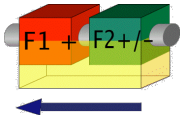
A promoter becomes functional due to organization of its elements


© 2004 by Genomatix Software GmbH

# Genomatix Functional organization of transcription



Physical vs functional TF binding sites




Transcriptional modules

Comparative genomics of promoters

**Understanding Promoter function Part II**

Regulatory networks

Promoter functions

Common TFBS vs frameworks



© 2004 by Genomatix Software GmbH

## Many genes have more than one physical promoter (sequence)

Multiple promoters can be responsible for multiple transcripts

- Best way to confirm multiple promoters is to look at their phylogeny
- ***ElDorado*** contains whole genome sequences for several mammals
- Comparative genomics of promoters is a unique feature of ***ElDorado***
- This allows to collect support for different promoter sets
- One promoter set contains all promoters for one ortholog transcript set

Phylogenetic conservation of promoters is the best safeguard against artifacts

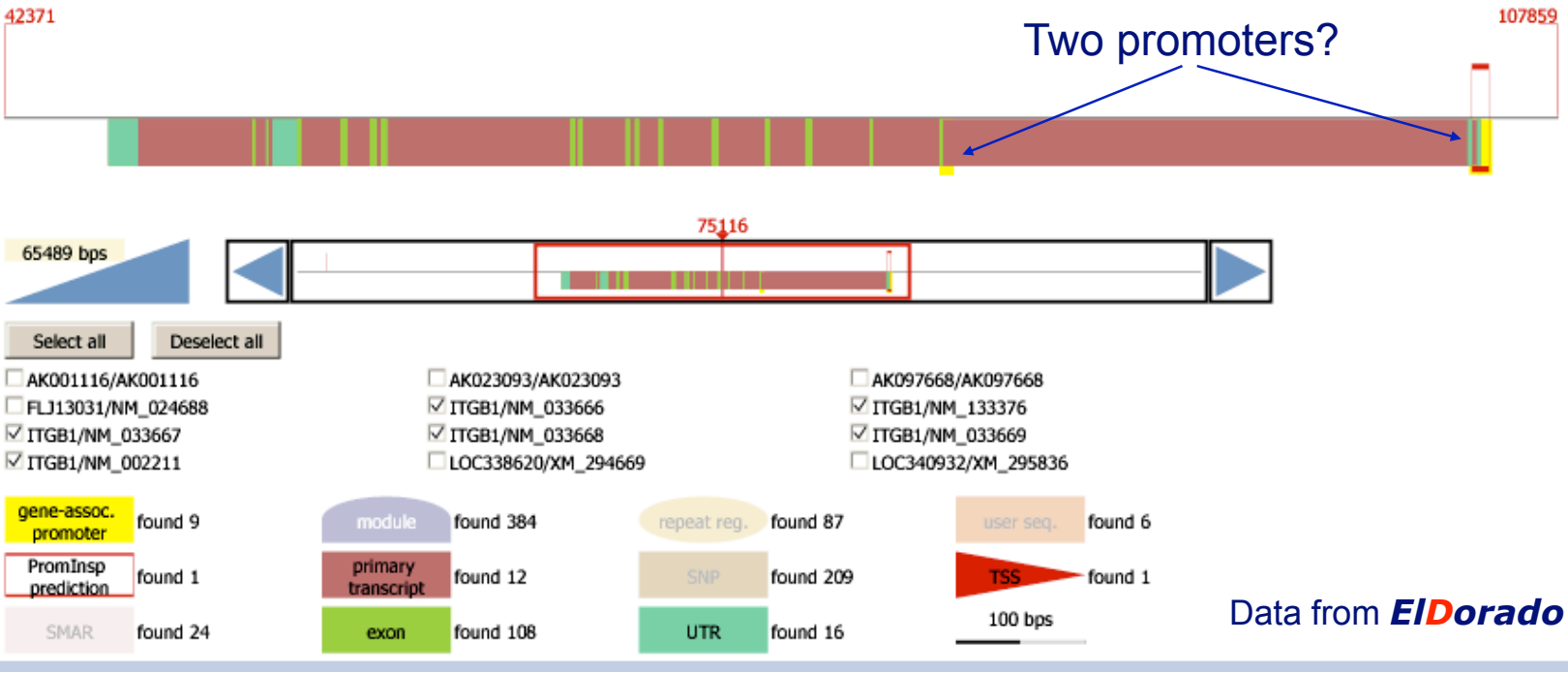

© 2004 by Genomatix Software GmbH

# Genomatix Part II: Comparative promoter genomics

## Genomic map of the beta1 integrin ITGB1

**Your input:** NM_033669(=ITGB1, CD29,FNRB,GPIIA,MDF2,MSK12,VLAB)
Your query was mapped to / found on NT_008705 (-) on chromosome 10 of **Homo sapiens**
NCBI human genome, build 33 (Apr 2003)
Extracted fragment: NT_008705 between 15117916 and 15277637 (159722 bps).

**Overview of extracted fragment**

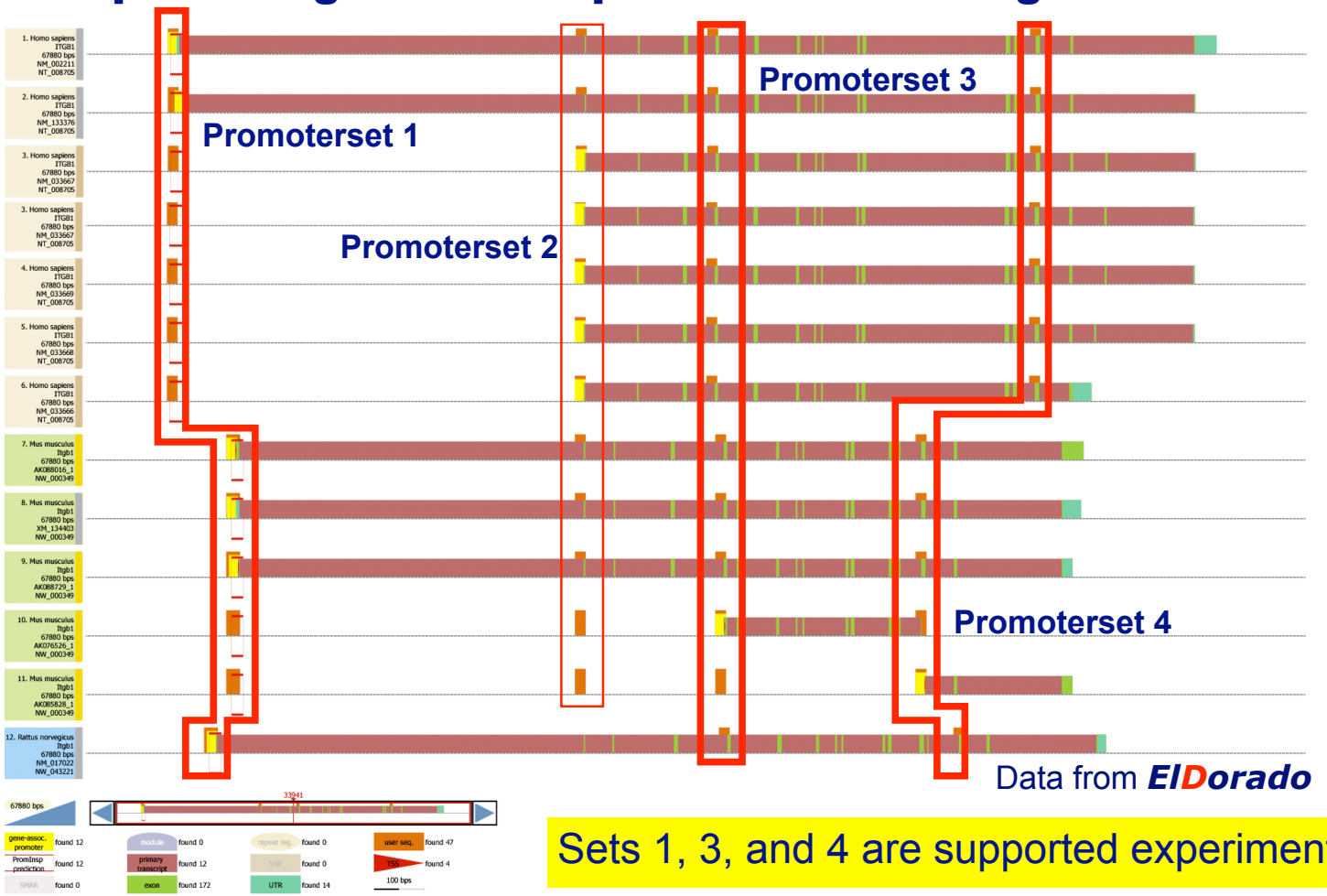




© 2004 by Genomatix Software GmbH

# Genomatix Part II: Comparative promoter genomics

## Comparative genomic map of the beta1 integrin ITGB1






© 2004 by Genomatix Software GmbH

## Why do genes have more than one promoter?

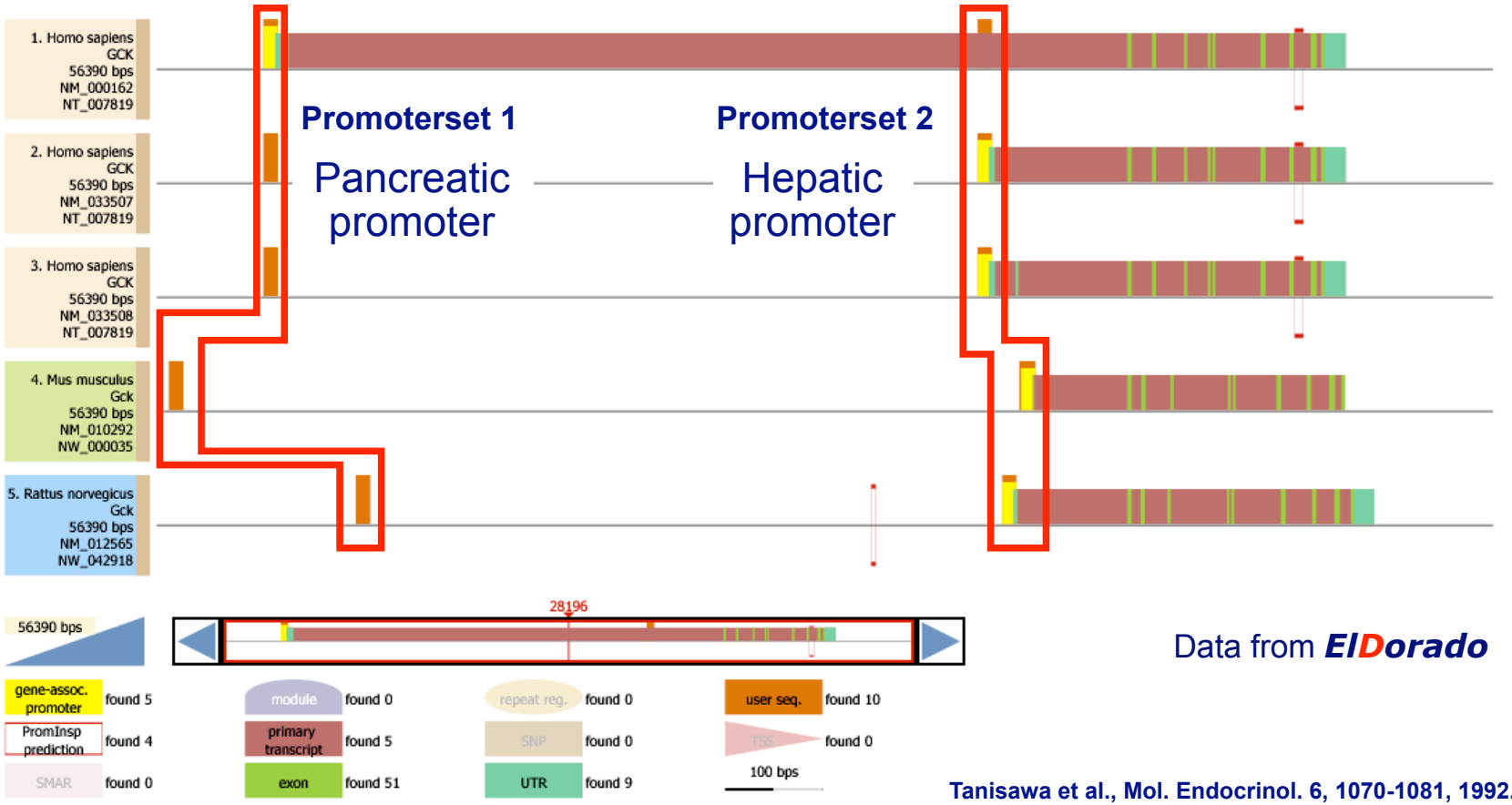
A significant portion of mammalian genes has more than one promoter

- Two promoters can direct expression independently in two tissues/cells
- Each promoter can have a much simpler structure
- Not all functional promoters are necessarily evolutionary conserved
- New expression patterns can be acquired by recombination
- Alternative transcripts can be realized in the most economic way

Genes with multiple promoters are common, not exceptions!


© 2004 by Genomatix Software GmbH

## Comparative genomic map of the Glucokinase GCK




Data from ***ElDorado***

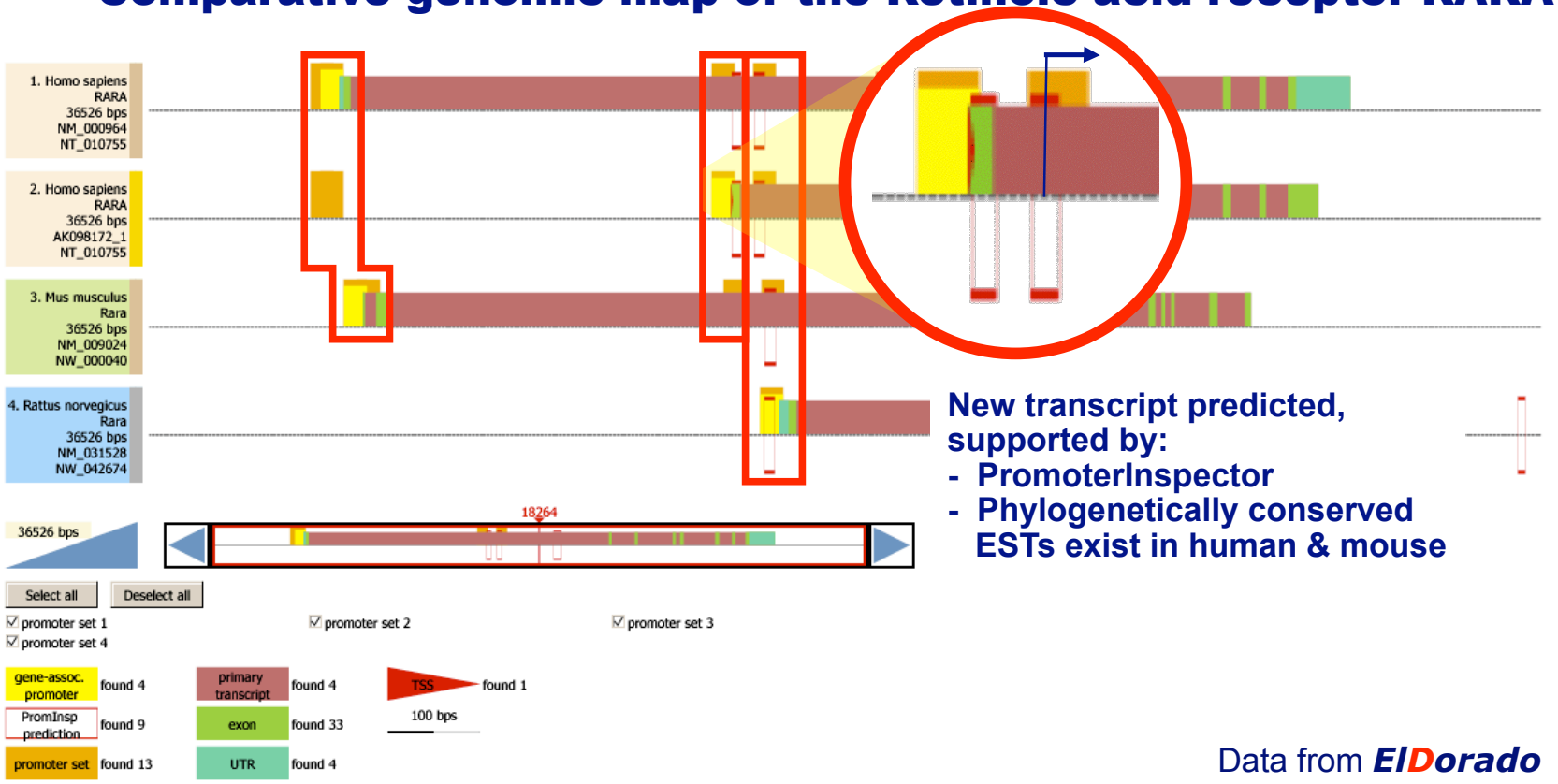
Tanisawa et al., Mol. Endocrinol. 6, 1070-1081, 1992.



© 2004 by Genomatix Software GmbH

# Genomatix Part II: Comparative promoter genomics

## Comparative genomic map of the Retinoic acid receptor RARA




**New transcript predicted, supported by:**
- **PromoterInspector**
- **Phylogenetically conserved ESTs exist in human & mouse**

Data from ElDorado

This gene appears to feature species specific transcripts


© 2004 by Genomatix Software GmbH

## Summary

Phylogenetic comparison is an important aid in promoter analysis

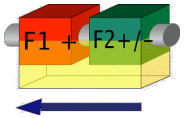
- Promoter region location is often phylogenetically conserved
- The sequence of promoter regions is often not conserved
- Direct alignment of orthologous promoter regions often does not work
- Genes can have multiple promoter regions

Promoters are conserved mainly in their organization, not their sequence

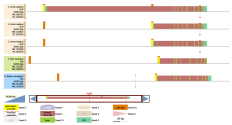

© 2004 by Genomatix Software GmbH

# Genomatix Functional organization of transcription

Physical vs functional
TF binding sites




Transcriptional
modules



Comparative
genomics
of promoters

**Understanding Promoter function Part II**

Regulatory networks

Promoter functions

Common TFBS vs frameworks



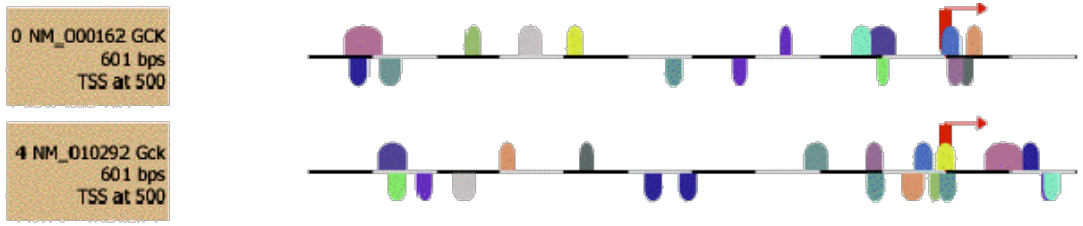
© 2004 by Genomatix Software GmbH

# Genomatix Part II: Common TFBS vs frameworks

## Comparison is only useful if functional relationship is clear

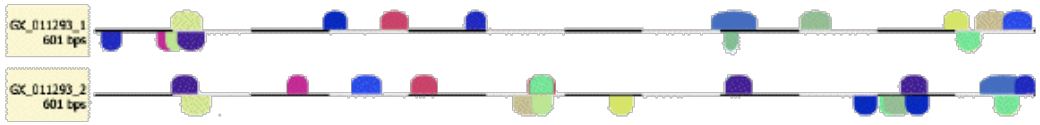
Orthologous GCK promoters

**12 common TFBs**



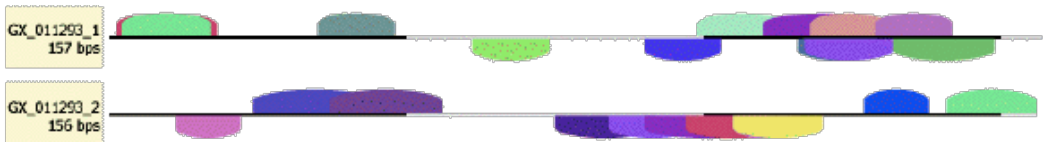

Unrelated GCK promoters

**14 common TFBs**




Coding exons GCK / MYLC2A

**20 common TFBs**




The orthologous promoters share the least TFBS

© 2004 by Genomatix Software GmbH

# Part II: Common TFBS vs frameworks

## Comparison is only useful if functional relationship is clear

Common TFBSs do not prove any functional relationship

- Common TFBSs are not suitable to detect promoters in other species
- First make sure promoter/enhancer sequences will be analyzed
- Make sure the set of sequences shares some transcriptional function
- Go for frameworks of TFBSs rather than individually shared TFBSs

Combine independent data: Shared transcriptional function & common TFBSs


© 2004 by Genomatix Software GmbH

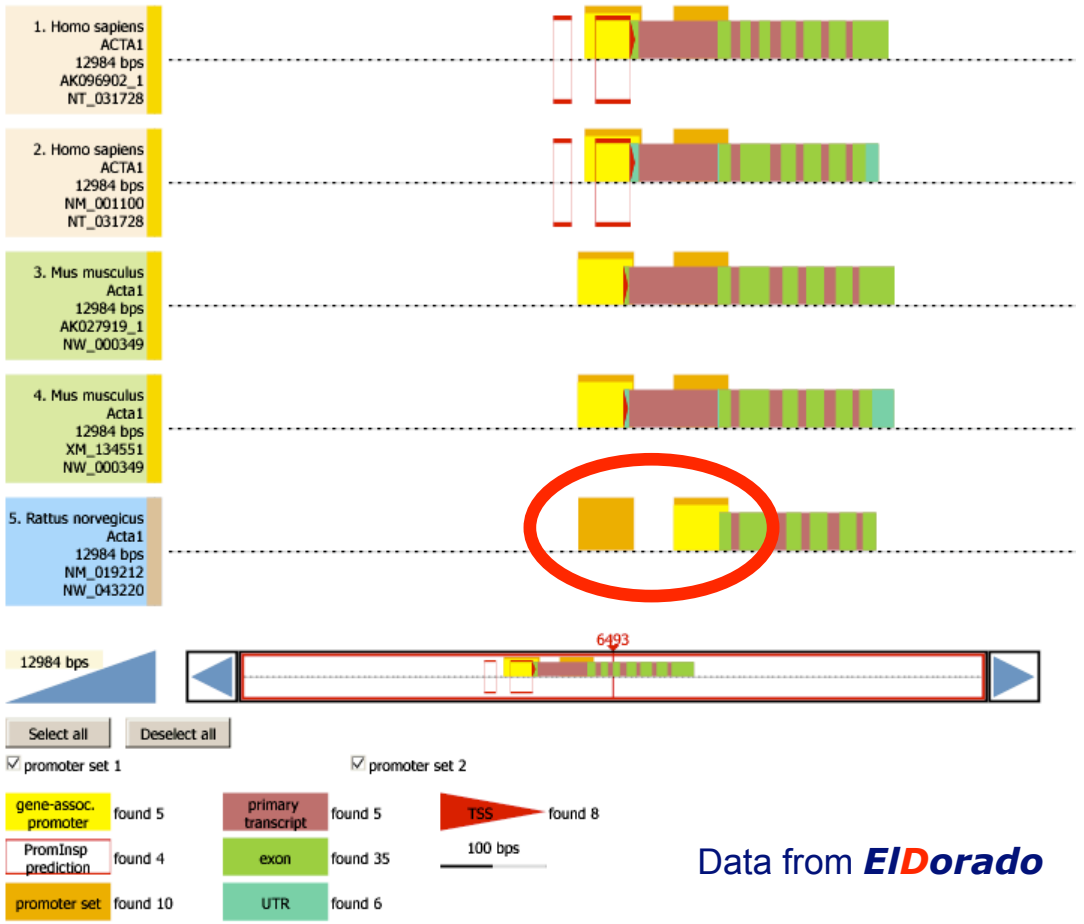
# Part II: Common TFBS vs frameworks

## Comparative genomics of the Actin A1 promoter

Human

Mouse

Rat



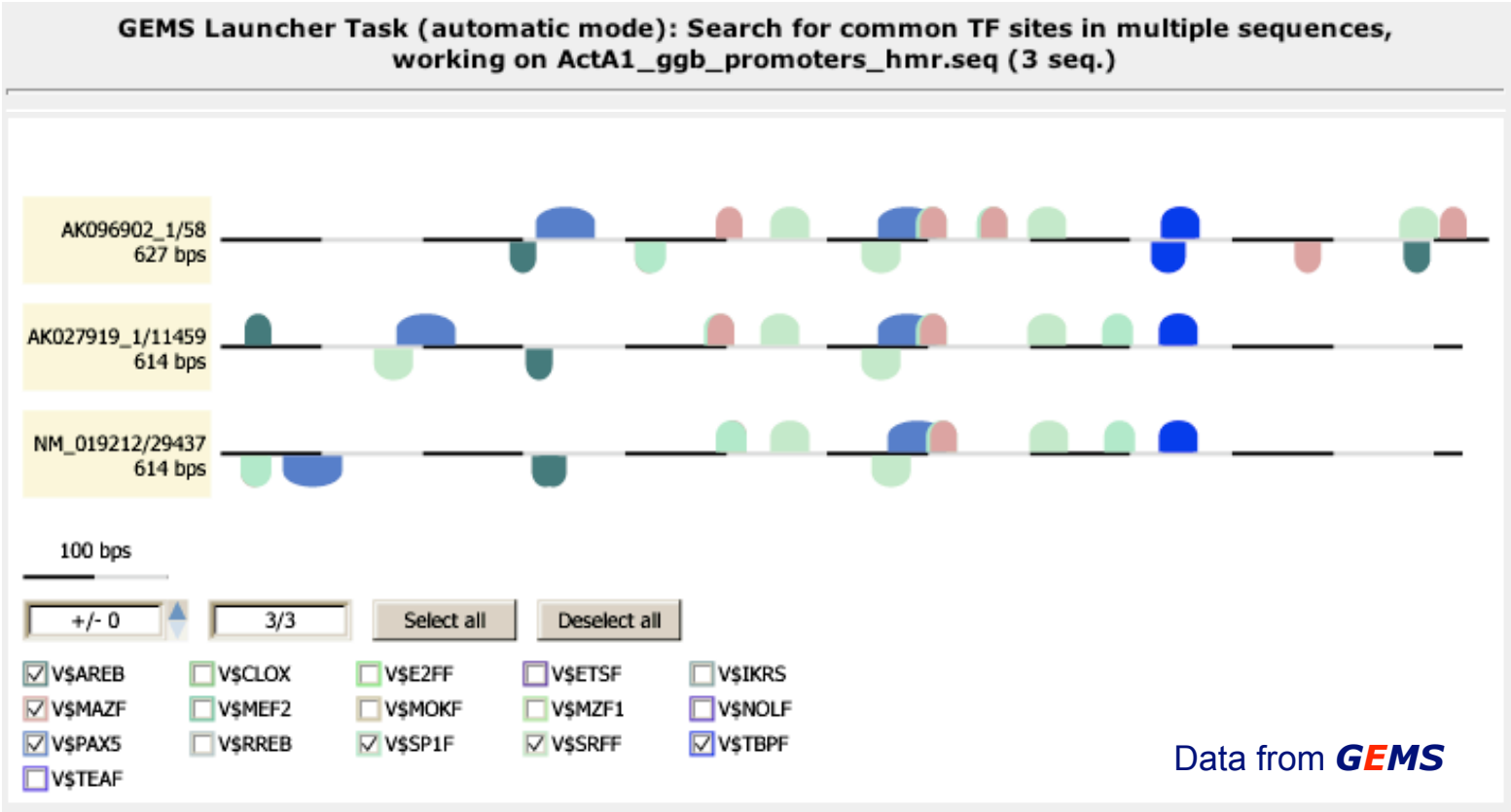

Data from ***ElDorado***



© 2004 by Genomatix Software GmbH

# Part II: Common TFBS vs frameworks

## Comparative genomics of the Actin A1 promoter

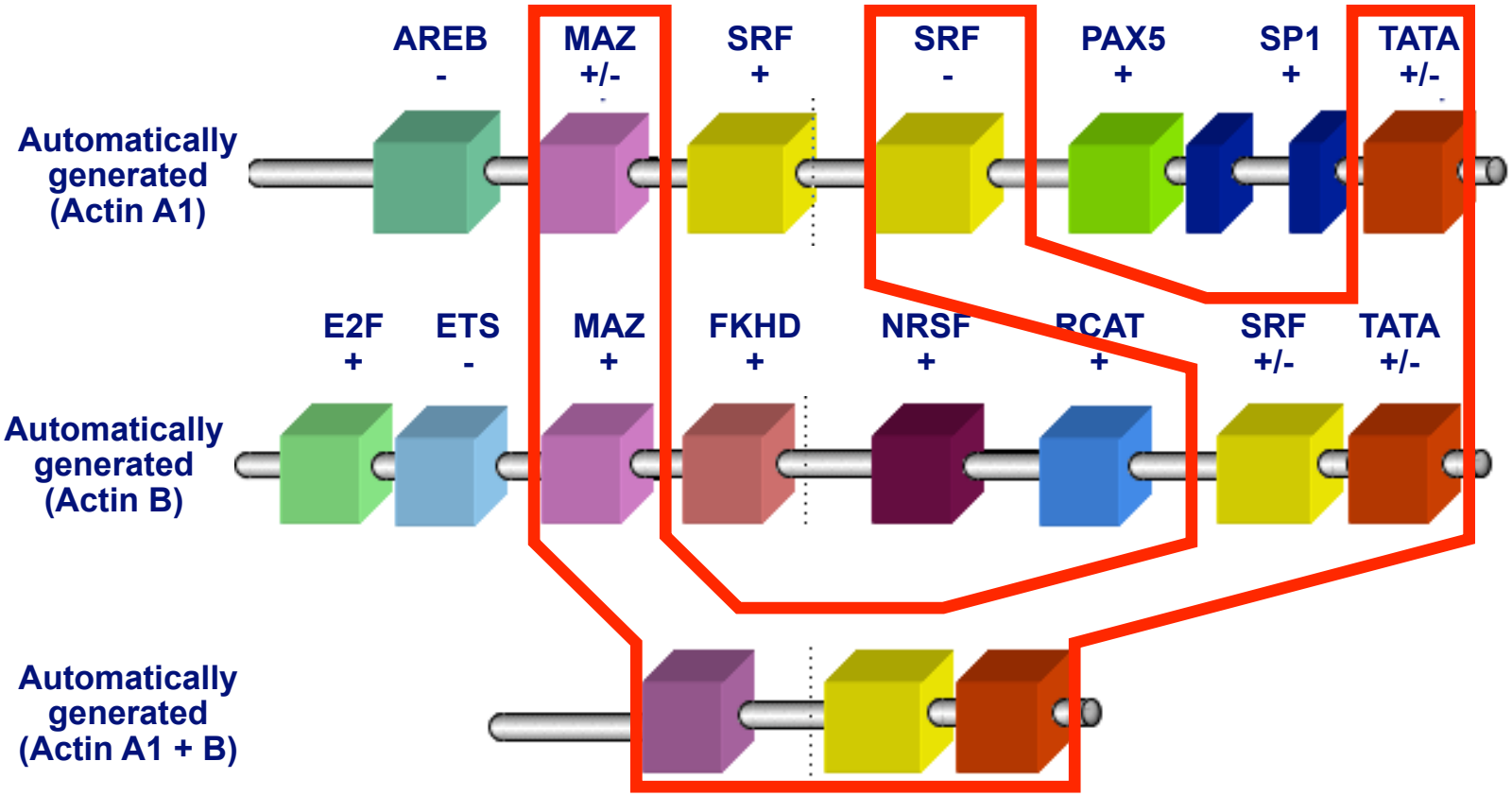





© 2004 by Genomatix Software GmbH

# Part II: Common TFBS vs frameworks

## Actin A1 framework versus actin B framework




Promoter frameworks depend on the set of sequences used for analysis



© 2004 by Genomatix Software GmbH

# Genomatix Part II: Common TFBS vs frameworks

## Actin A1 promoters

- All modules are ex
- A single module is

| Sequence |
|---|
| **096902_1_promoter** (1 - 601), fil<br>G3-actinA1_g_promoters_hm.seq<br>**ACTA1/AK096902_1_promoter [AK0969**<br>**278542-279142=601bps) [oligo_capped A** |
| **027919_1_promoter** (1 - 601), file<br>_g_promoters_hm.seq<br>**AK027919_1_promoter [AK0279**<br>**51716056-51716656=601bps) [oligo_cap**<br>**TSS=500** |
| **043220_193584** (1 - 601), file G3-actin<br>**V_043220_193584 (?), sequence len**<br>**complement, NW_043220_193584 (NW_0**<br>**193584 to 194184, original sequence lengt**<br>**norvegicus chromosome 19 WGS supercont** |

Human

Mouse

Rat

**Model YY1F_SRFF**

| | |
|---|---|
| Model Name | YY1F_SRFF_02 (YY1F - SRFF n |

**Model**

| | Element type | Name | Strand | Para... | ...ment |
|---|---|---|---|---|---|
| 1 | Matrix | V$YY1F | (-) | Min. core sim.: 0.980<br>Min. matrix sim.: 0.769 | 0 to 4 bp |
| 2 | Matrix | V$SRFF | (+) | Min. core sim.: 0.952<br>Min. matrix sim.: 0.938 | --- |

**Total length:** 0 - 4 bp

**Optimized model threshold:** 80 %

**Origin**

| | |
|---|---|
| **Reference** | Gualberto et al., Mol. Cell. Biol. 12, 4209-4214, 1992 (MEDLINE: 1508214) |
| **Gene** | Chicken skeletal alpha-actin gene. |
| **Accession no.** | M13631; |

**Function**

The factors in this module antagonize in the regulation of chicken skeletal alpha actin gene.

**Quality Assessment:**

0.010 matches per 10,000 bps in human sequences

Data from **GEMS**

**Min. matrix sim: 0.769**

**Opt. Matrix sim.: 0.840**

The YY1 - SRF module was detected and defined in a chicken actin gene



© 2004 by Genomatix Software GmbH

# Part II: Common TFBS vs frameworks

## Summary

Comparative genomics can help elucidate functional promoter structures

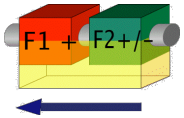
- Common TFBSs alone do prove any biological function
- Only compare orthologous or functionally related promoters
- Functional relationship is seen in conserved TFBS frameworks
- Conserved transcriptional modules suggest conserved function

Phylogenetically conserved promoter structures indicate functional conservation

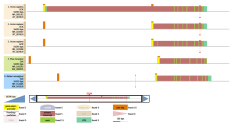

© 2004 by Genomatix Software GmbH

# Genomatix Functional organization of transcription

Physical vs functional
TF binding sites



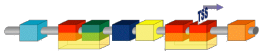

Transcriptional
modules



Comparative
genomics
of promoters

**Understanding
Promoter function
Part II**

Regulatory
networks



Common TFBS
vs frameworks

Promoter
functions



© 2004 by Genomatix Software GmbH

# Part II: Promoter functions

## The organization of promoters is functionally restricted

- Promoters must integrate several signals into one output
- Promoters must support several distinct 3-D protein complexes
- Promoters must correctly function in embryonal development
- Promoters must encode distinct functionalities in several somatic tissues
- All this must be achieved with a limited number of elements

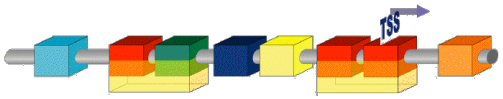
The sequence of a promoter region is fixed in all tissues, its function changes!


© 2004 by Genomatix Software GmbH

# Part II: Promoter functions

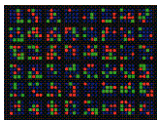
## Advantages and limitations of *in silico* promoter analysis

- Regulatory „thermodynamics“




  - Genomic promoter sequences are static
  - Every functional dimension is imprinted in this sequence
  - There is no information WHEN functions will be active

- Regulatory „kinetics“



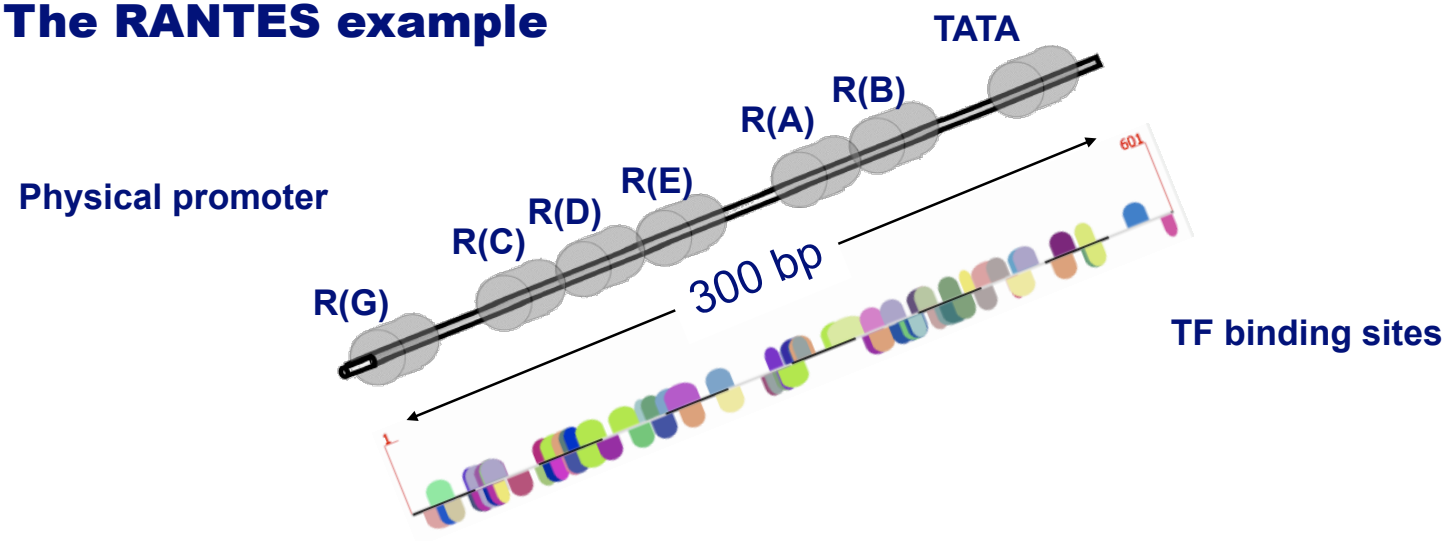
  - (Runon) expression arrays indicate WHEN and WHERE transcription is active
  - Expression arrays do NOT indicate WHY transcription is active

Understanding transcriptional regulation requires both approaches


© 2004 by Genomatix Software GmbH

# Part II: Promoter functions

## The RANTES example




- Seven binding regions (84 TF potential binding sites)
- Tissue specific regulation encoded in promoter
- Experimentally well characterized

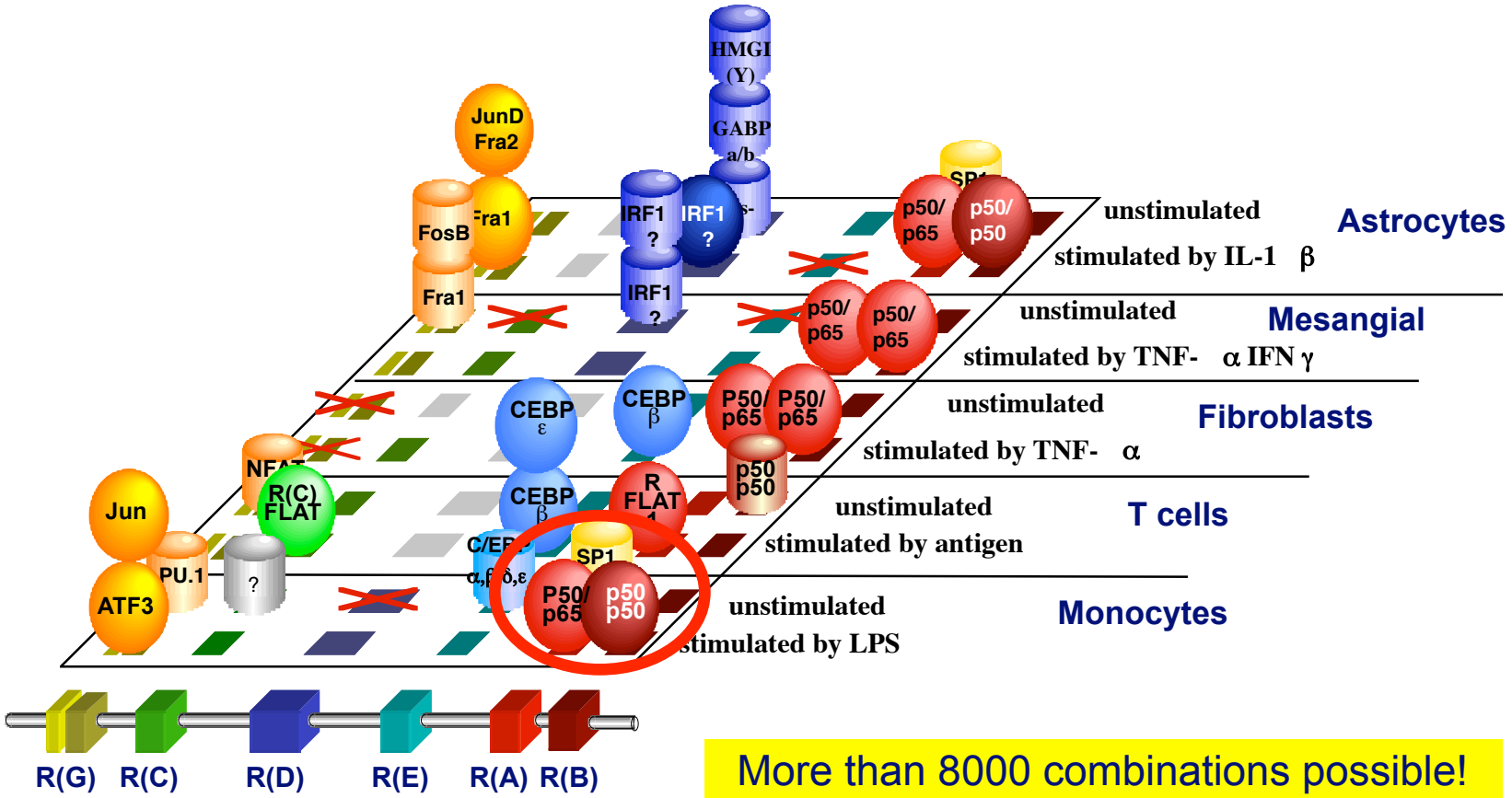
Regulation of the RANTES promoter is very complex


© 2004 by Genomatix Software GmbH

# Part II: Promoter functions

## The RANTES example

- Six binding sites
- Five cell lines




More than 8000 combinations possible!



© 2004 by Genomatix Software GmbH

# Part II: Promoter functions

## The RANTES A region NFkB and SP1 binding?

**Inspecting sequence RANTES_NFkB_regionAB (1 - 57):**

| Family/matrix | Further Information | Opt. | Position from - to | Position anchor | Str. | Core sim. | Matrix sim. | Sequence |
|---|---|---|---|---|---|---|---|---|
| [illegible] | muscle-specific Mt binding site | 0.90 | 4 - 12 | 8 | (+) | 1.000 | 0.922 | tgctATTTt |
| V$NFKB/NFKAPPAB65.01 | [illegible]-kappaB (p65) | 0.87 | 10 - 24 | 17 | (-) | 1.000 | 0.984 | aggggagtTTCCaaa |
| V$SP1F/BTEB3.01 | Basic transcription element (BTE) binding protein, BTEB3, FKLF-2 | 0.97 | 12 - 26 | 19 | (-) | 1.000 | 0.956 | t |
| V$NOLF/OLF1.01 | Olfactory neuron-specific factor | 0.82 | 13 - 35 | 24 | (+) | 1.000 | 0.858 | ggaaacTCCCcttaggggatgcc |
| V$NFKB/NFKAPPAB50.01 | NF-kappaB (p50) | 0.83 | 25 - 39 | 32 | (+) | 1.000 | 0.969 | taGGGGatgcccctc |
| V$CMYB/CMYB.01 | c-Myb, important in hematopoesis, cellular equivalent to avian myoblastosis virus oncogene v-myb | 0.99 | 36 - 44 | 40 | (-) | 1.000 | 0.990 | caGTTGagg |
| V$PAX6/PAX6.02 | PAX6 paired domain and homeodomain are required for binding to this site | 0.89 | 36 - 54 | 45 | (-) | 1.000 | 0.907 | tttatagggCCAGttgagg |

A region

7 matches found.

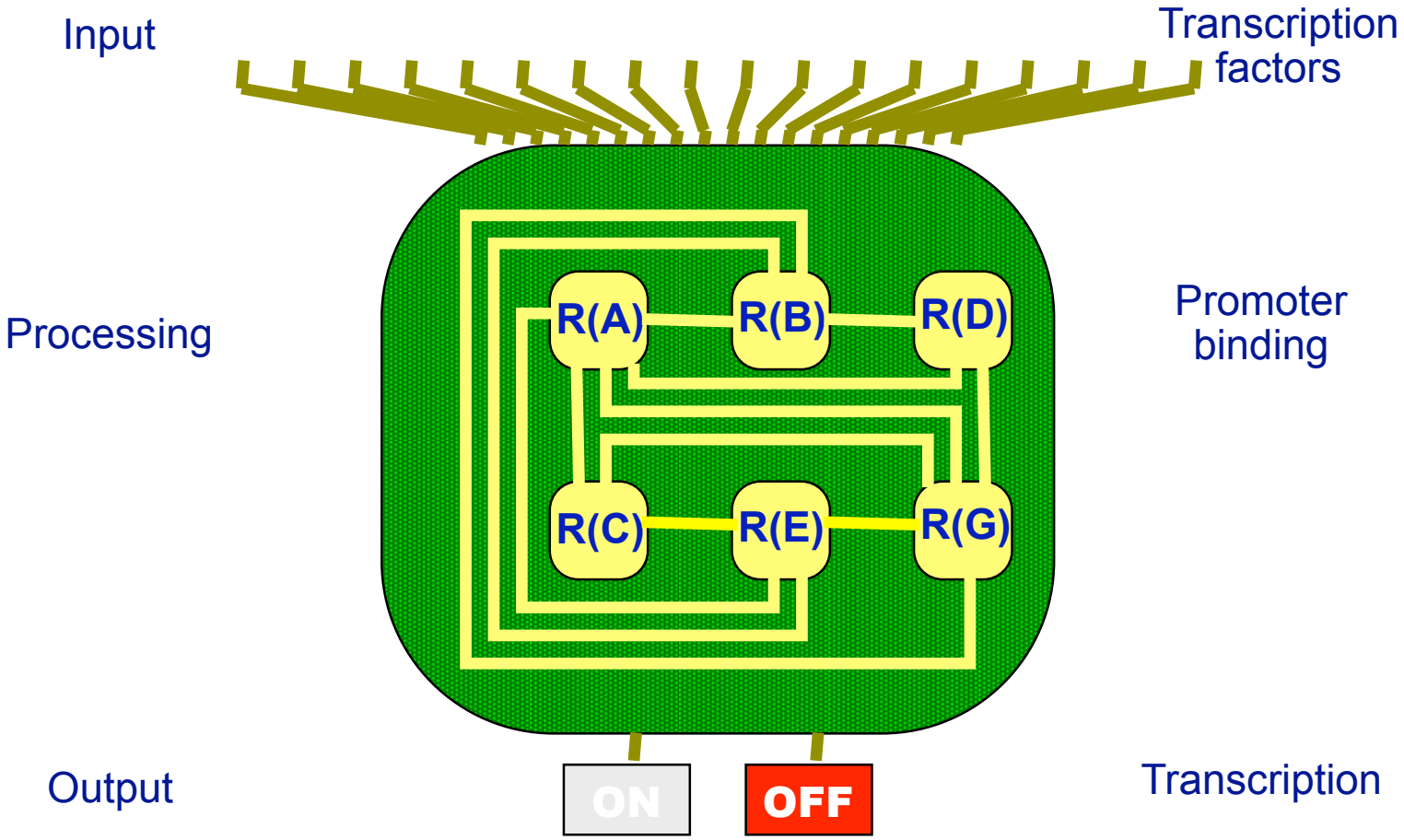

***MatInspector*** correctly predicted the overlapping binding sites



© 2004 by Genomatix Software GmbH

# Promoter Part II: Promoter functions

## The RANTES promoter circuits

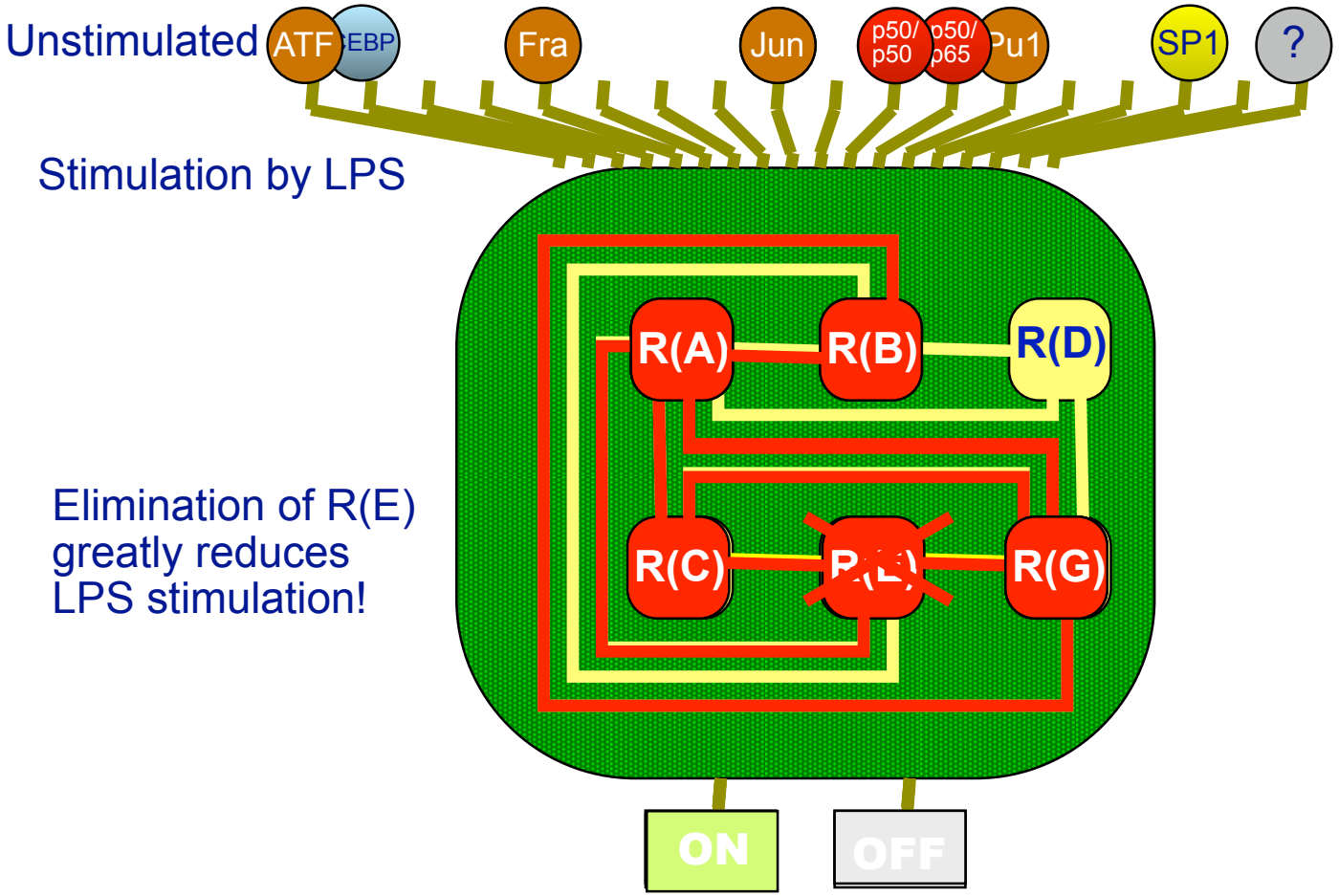





© 2004 by Genomatix Software GmbH

# Part II: Promoter functions

## The RANTES promoter: the monocyte program

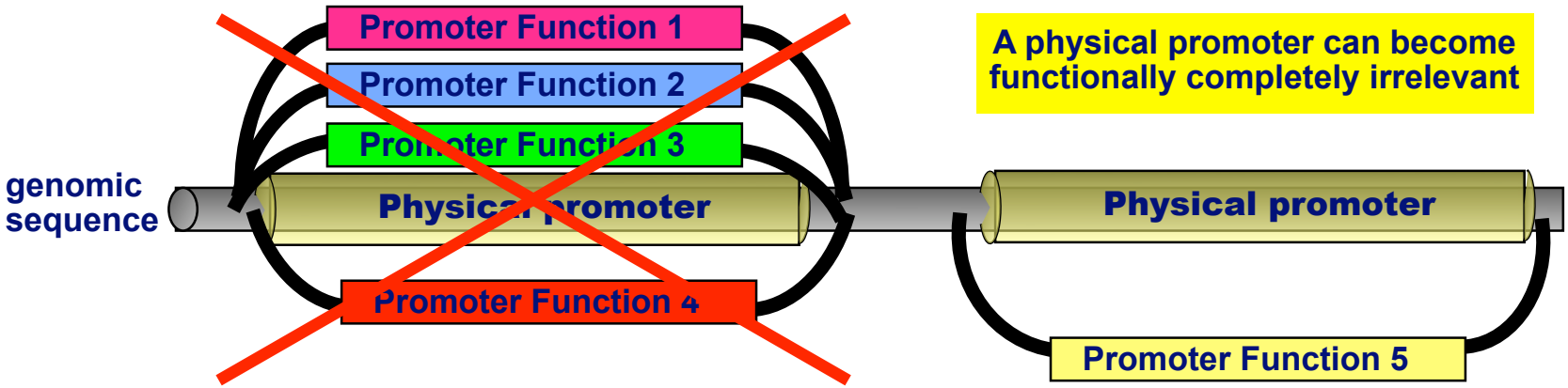





© 2004 by Genomatix Software GmbH

# Part II: Promoter functions

## Physical promoter versus functional promoter complexity




- Physical promoter region
  - Detectable in genomic sequence
  - Fixed sequence / many TF sites
  - Fixed genomic location
  - Variable structure (chromatin)
- Promoter function
  - Not detectable in genomic sequence
  - Variable organization (multiple functions)
  - Multiple genomic locations
  - Cell-/Tissue specific functions

Drug action or disease related questions require analysis of promoter functions


© 2004 by Genomatix Software GmbH

# Part II: Promoter functions

## Summary

Promoter regions can fulfill many distinct promoter functions

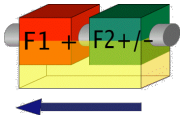
- All possible functions are encoded in the sequence of a promoter region
- The organization of a elements in promoter region resembles *hardware*
- Sets of TFs binding to the promoter region resemble *input*
- The TFs interactions on the promoter region resemble a *program*

The output of any promoter program is ON or OFF of transcriptional activity

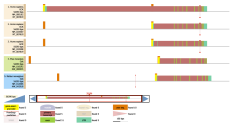

© 2004 by Genomatix Software GmbH

# Genomatix Functional organization of transcription

Physical vs functional
TF binding sites



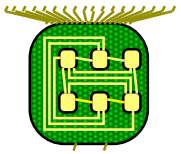

Transcriptional
modules



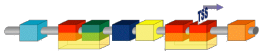
Comparative
genomics
of promoters

**Understanding Promoter function Part II**

Regulatory networks



Promoter
functions



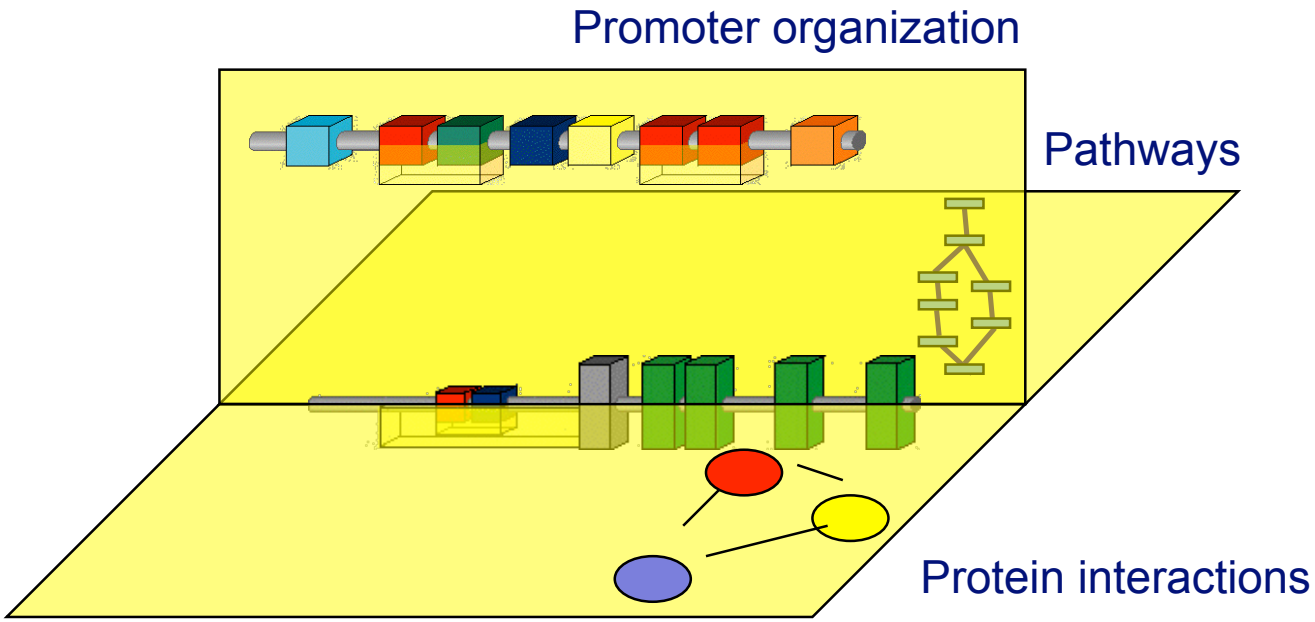
Common TFBS
vs frameworks



© 2004 by Genomatix Software GmbH

# Part II: Regulatory networks

## Promoter analysis reveals part of the functional context

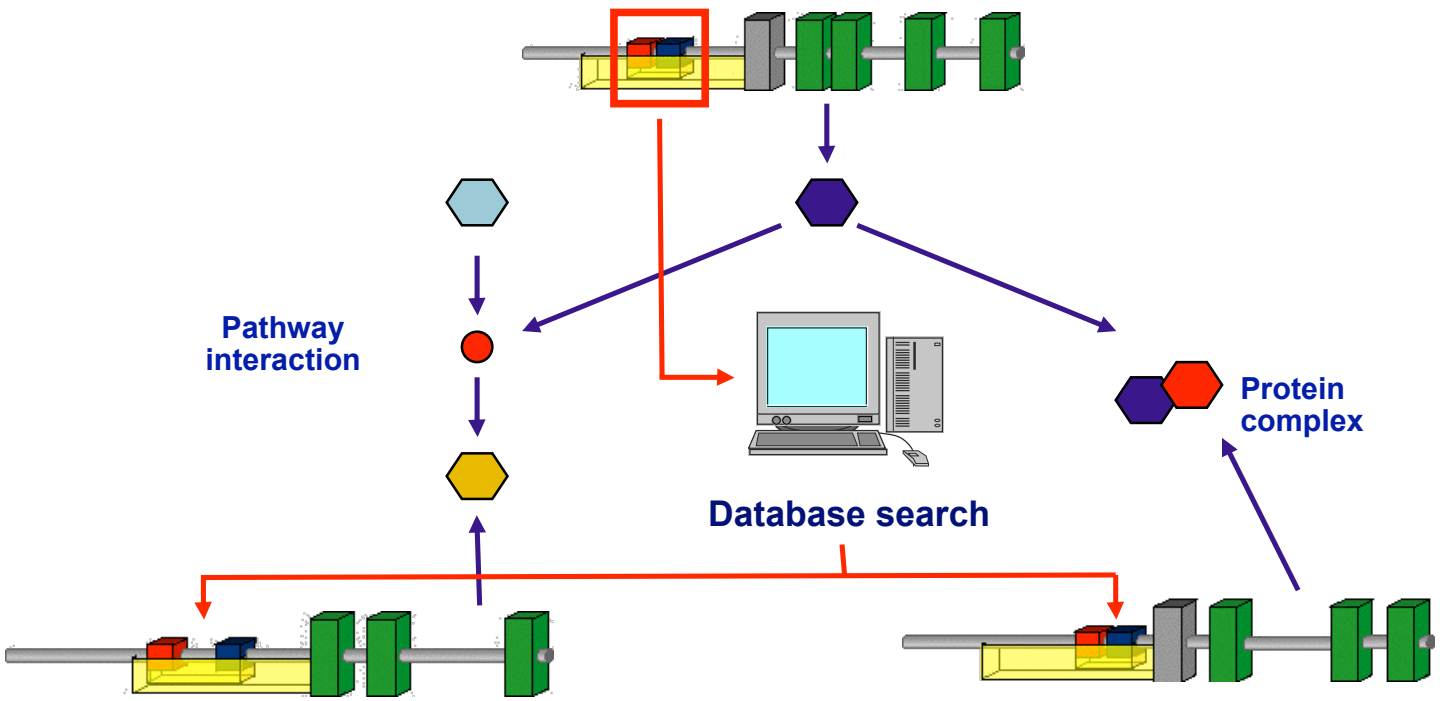



The context of expression is an essential part of the function of a gene


© 2004 by Genomatix Software GmbH

# Part II: Regulatory networks

## Tracing networks by promoter modules




Functional context can be derived from conditional regulatory context


© 2004 by Genomatix Software GmbH

# Part II: Regulatory networks

## Summary

Regulatory networks steer and link signaling and metabolic networks

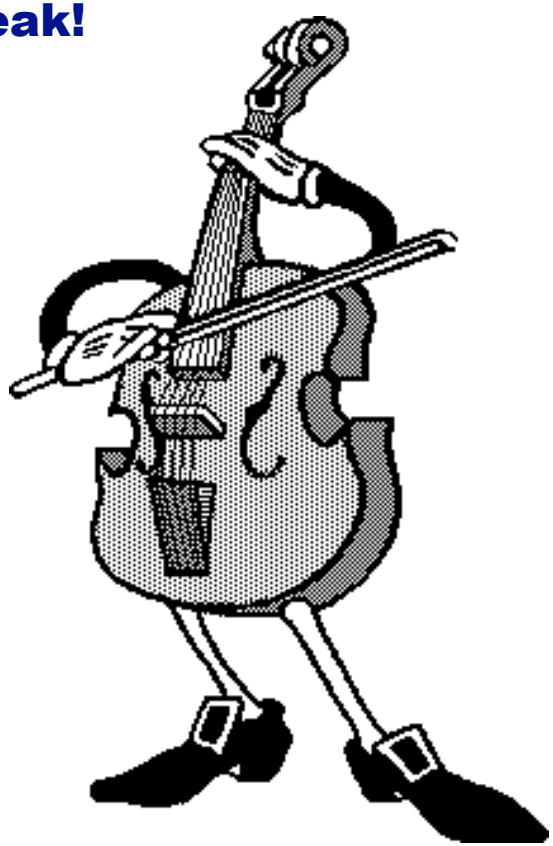
- Functional connections of genes are often reflected in the promoters
- Shared modules in promoters often indicate functional connection
- Functionally linked genes can be detected by promoter analysis
- Transcriptional networks mirror functional protein networks

Regulatory networks provide the basis for many physiological feedback loops


© 2004 by Genomatix Software GmbH

# End of part two

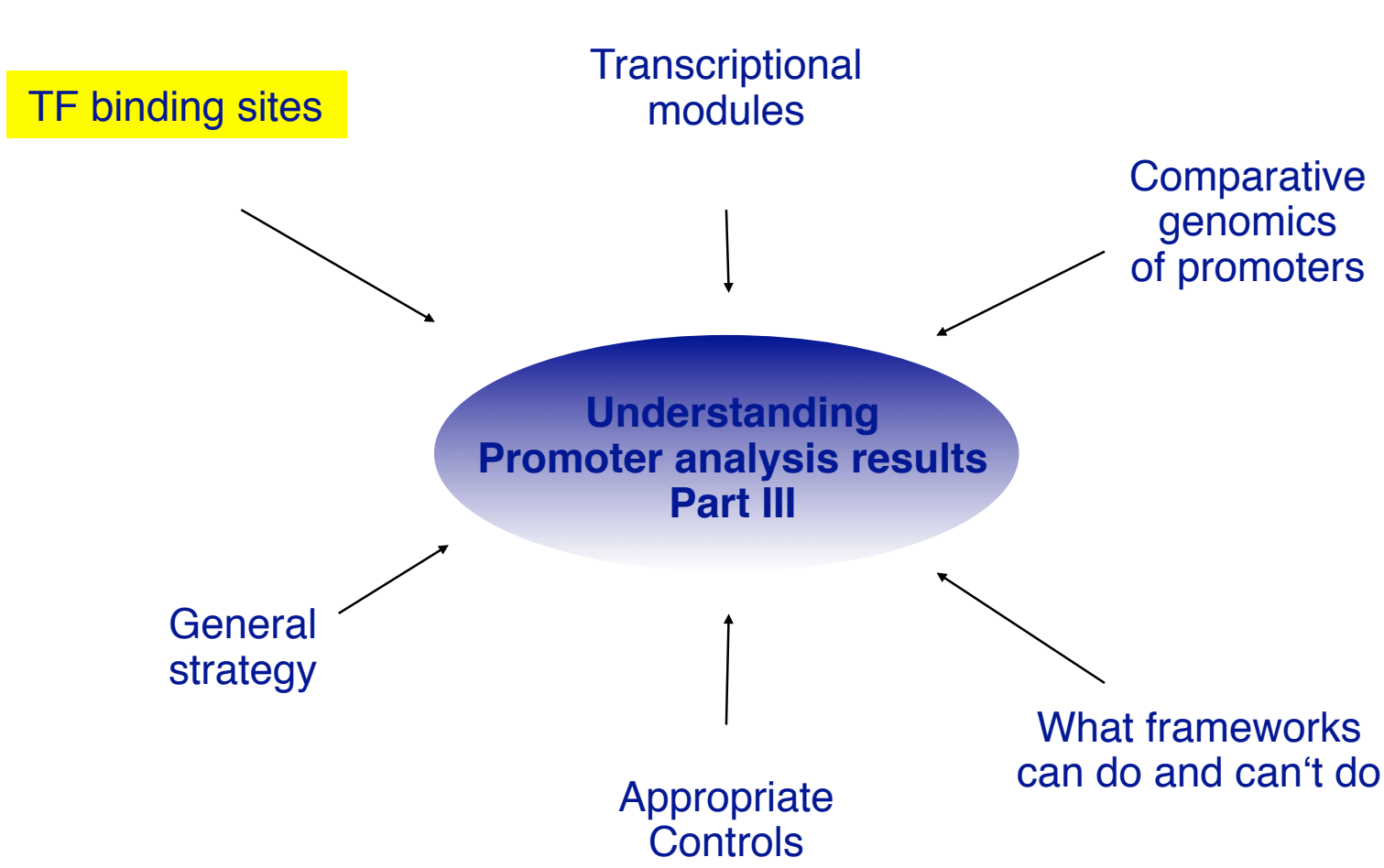
## Time for another break!





© 2004 by Genomatix Software GmbH

# Promoter analysis strategies





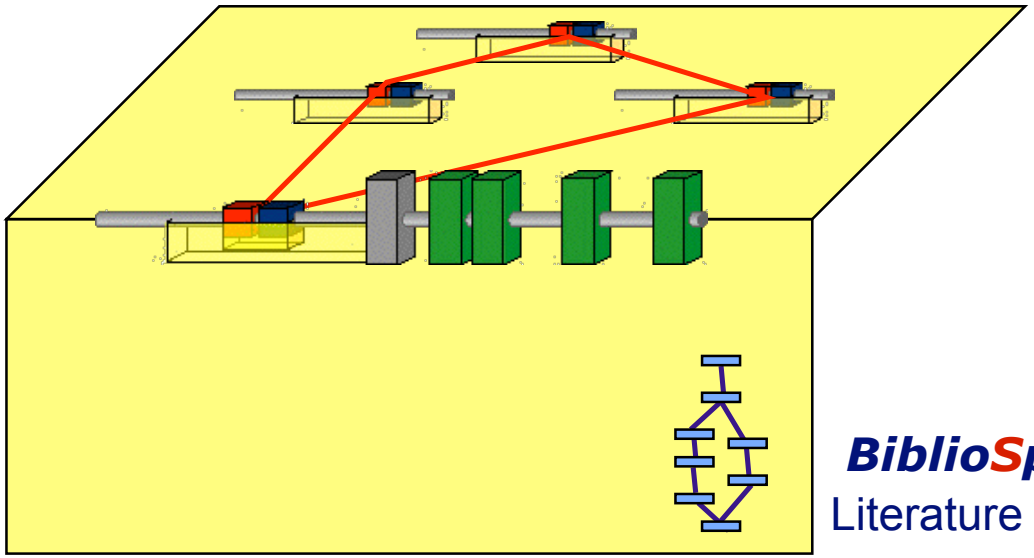
© 2004 by Genomatix Software GmbH

# Part III: TF binding sites

## Independent data sources enhance functional context analysis

Promoter network

***GEMS***



***BiblioSphere***

Literature analysis

Only independent data and algorithms allow a synergistic consensus approach

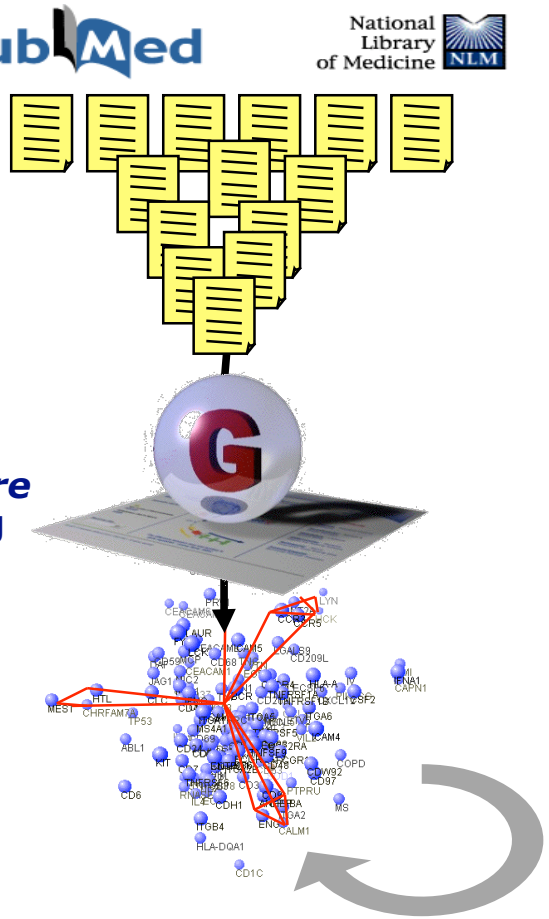


© 2004 by Genomatix Software GmbH

# Part II: Regulatory networks

## BiblioSphere

- 12 million abstracts
- Rapidly growing database
- About 1,100 papers/day




- Conserved structure (3D representation)
- Extracted correlations
- Directly linked to sequence analysis of genes&promoters

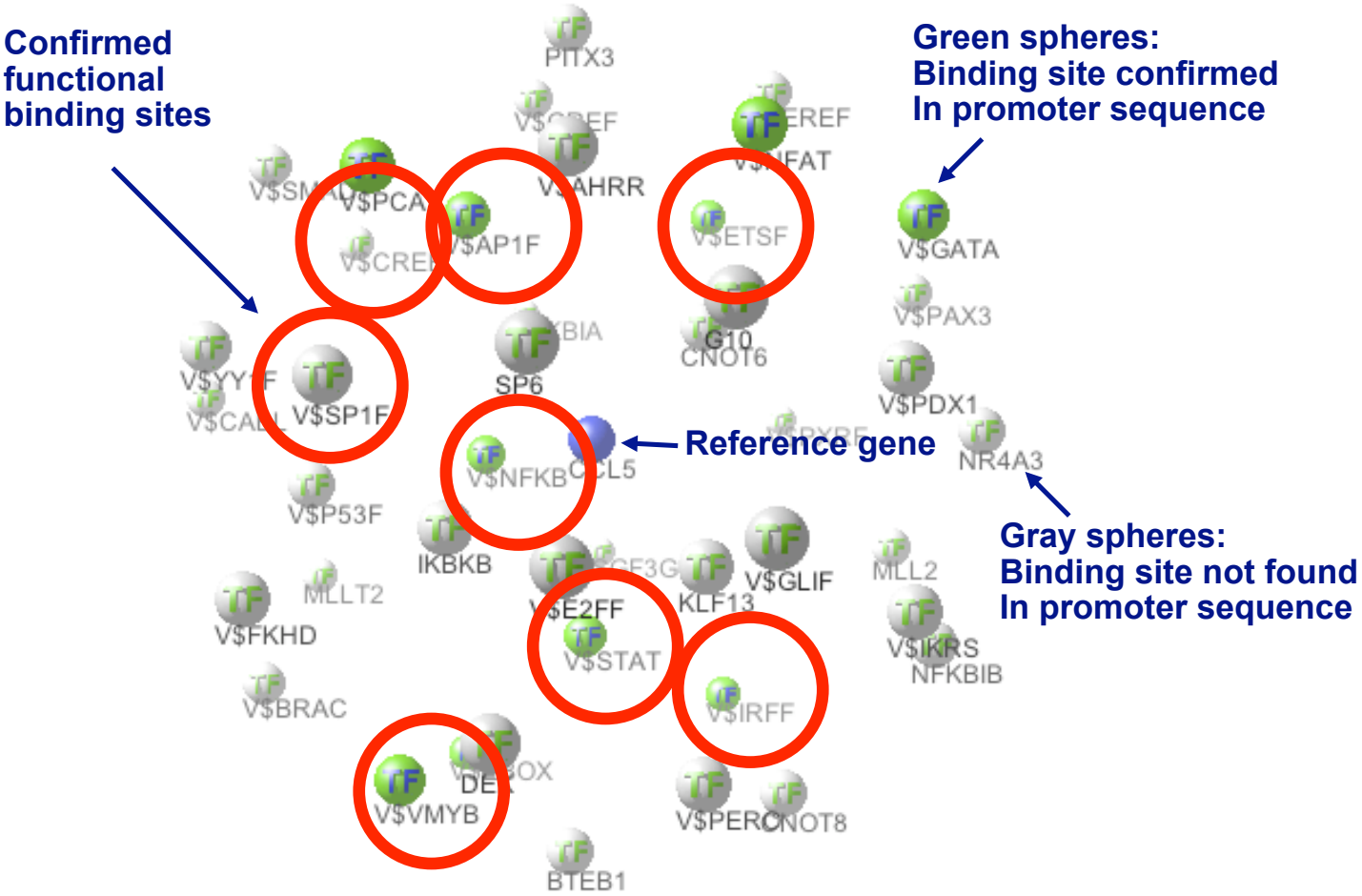
AUTOMATIC



© 2004 by Genomatix Software GmbH

# Part II: Regulatory networks

## Transcriptional context of RANTES identified by *BiblioSphere*

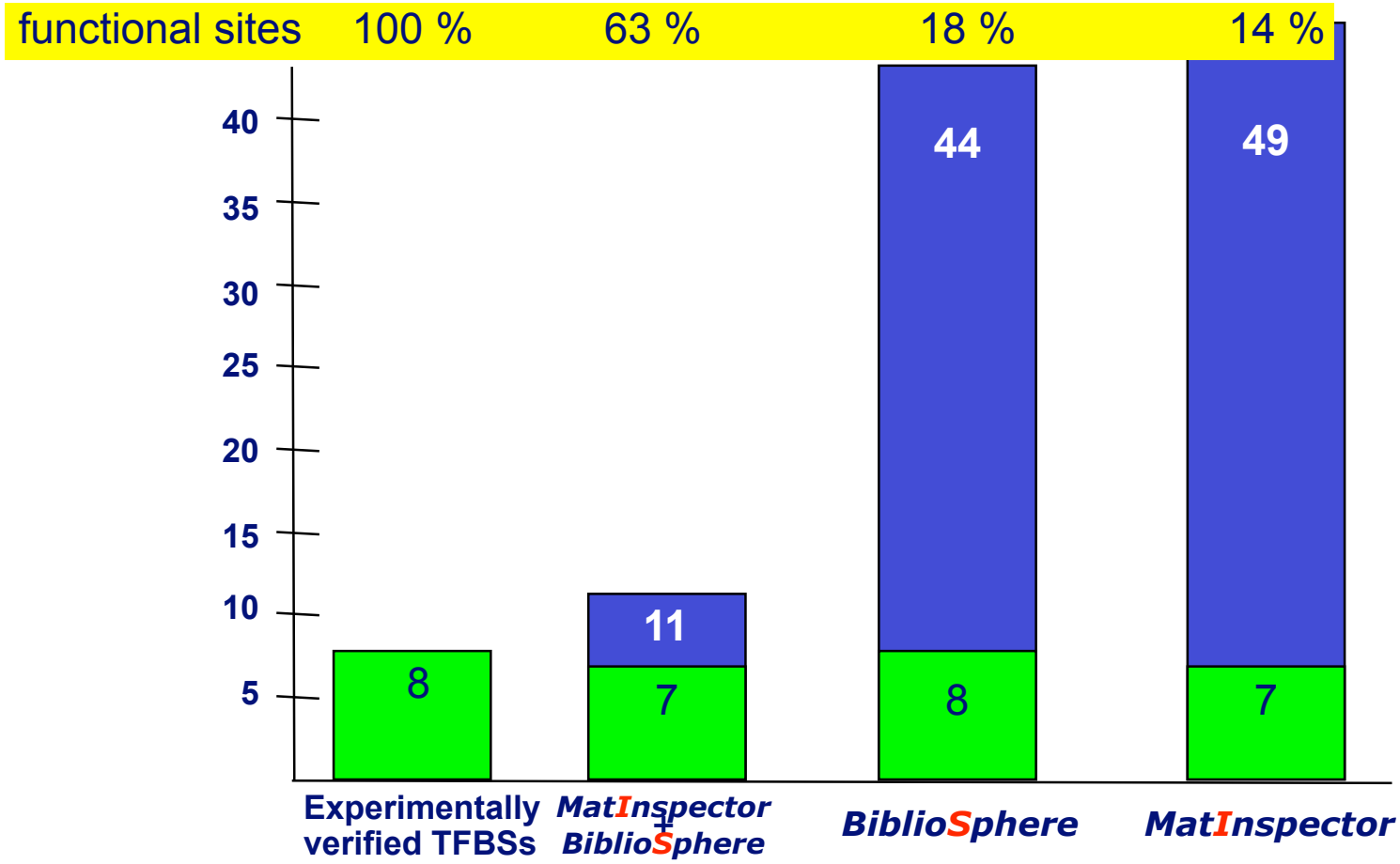





© 2004 by Genomatix Software GmbH

# Part II: Regulatory networks

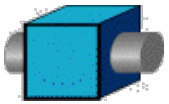
## Transcriptional context of RANTES - cross filtering






© 2004 by Genomatix Software GmbH

# Promoter analysis strategies



Transcription factor binding sites

Transcriptional modules

Comparative genomics of promoters

**Understanding Promoter analysis results Part III**

General strategy

Appropriate Controls

What frameworks can do and can‘t do

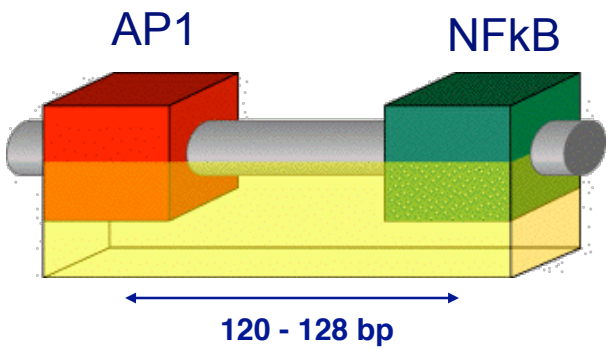


© 2004 by Genomatix Software GmbH

# Part III: Transcriptional modules

## Search with AP1-NFkB “module” in the EPD (2997 promoters)

- AP1-NFkB is a functionally verified promoter module




- Searching EPD with this module (***GEMS***) should find target genes!
- The human P53 gene is a known target (module derived from p53)


© 2004 by Genomatix Software GmbH

# Part III: Transcriptional modules

## Search with AP1-NFkB "module" in the EPD (2997 promoters)

- A total of 6 matches was found

| Sequence | Model Name | Position | Strand | Model Score | Select Match |
|---|---|---|---|---|---|
| **HS_TP53** [EP1122.] (1 - 600) **[DNA] Hs p53; range -499 to 100.** | AP1_NFkB_submodel | 424 - 566 | (+) | 95.7 % | |
| **MM_MHCD** [EP11156] (1 - 600) **[DNA] Mm MHCI H-2L^d; range -499 to 100.** | AP1_NFkB_submodel | 206 - 344 | (+) | 95.4 % | |
| **HS_ALAS1** [EP73282] (1 - 600) **[DNA] Hs ALAS1; range -499 to 100.** | AP1_NFkB_submodel | 466 - 324 | (-) | 97.1 % | |
| **SV40_TA_1** [EP07166] (1 - 600) **[DNA] SV40 T/t late P1; range -499 to 100.** | AP1_NFkB_submodel | 492 - 353 | (-) | 99.4 % | |
| **SV40_TA_2** [EP07164] (1 - 600) **[DNA] SV40 T/t early P2; range -499 to 100.** | AP1_NFkB_submodel | 492 - 353 | (-) | 99.4 % | |
| **SV40_COAL** [EP07168] (1 - 600) **[DNA] SV40 late; range -499 to 100.** | AP1_NFkB_submodel | 284 - 423 | (+) | 99.4 % | |

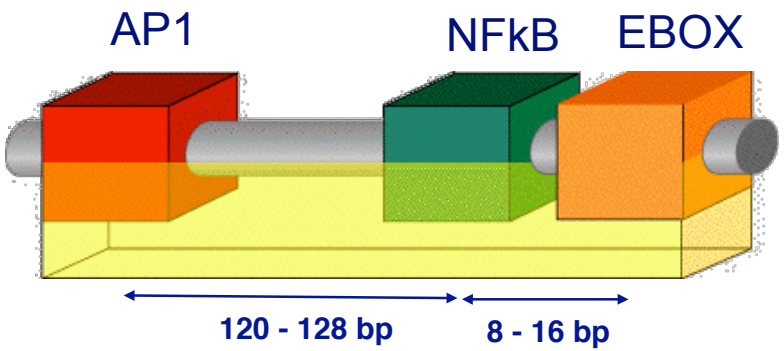
However, this does not necessarily indicate six target genes



© 2004 by Genomatix Software GmbH

# Part III: Transcriptional modules

## AP1-NFkB is only a submodel of a functional module




- The complete AP1-NFkB-EBOX module finds 1 match in EPD

| Sequence | Model Name | Position | Strand | Model Score | Select Match |
|---|---|---|---|---|---|
| HS_TP53 [EP11223] (1 - 600) [DNA] Hs p53, range -499 to 100. | AP1F_NFKB_EBOX_01 | 424 - 578 | (+) | 94.9 % | |

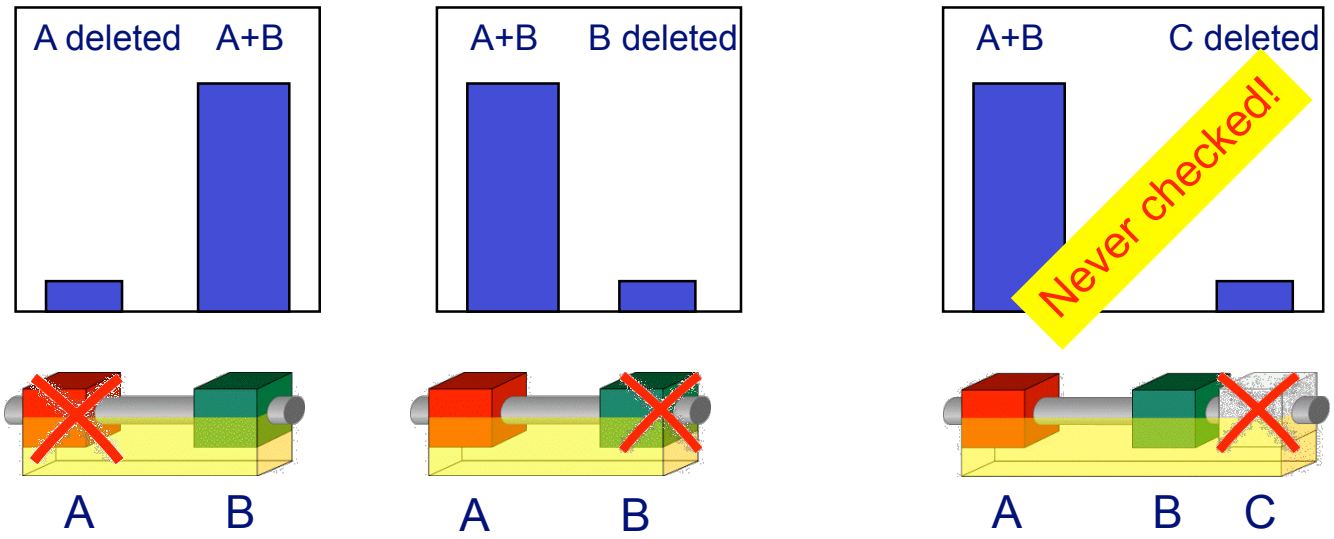
Only the complete functional module is target gene specific!


© 2004 by Genomatix Software GmbH

# Part III: Transcriptional modules

## Why can verified modules still be incomplete?

- Modules are usually verified by mutual deletion experiments
- Loss of function upon deletion of either site proves module



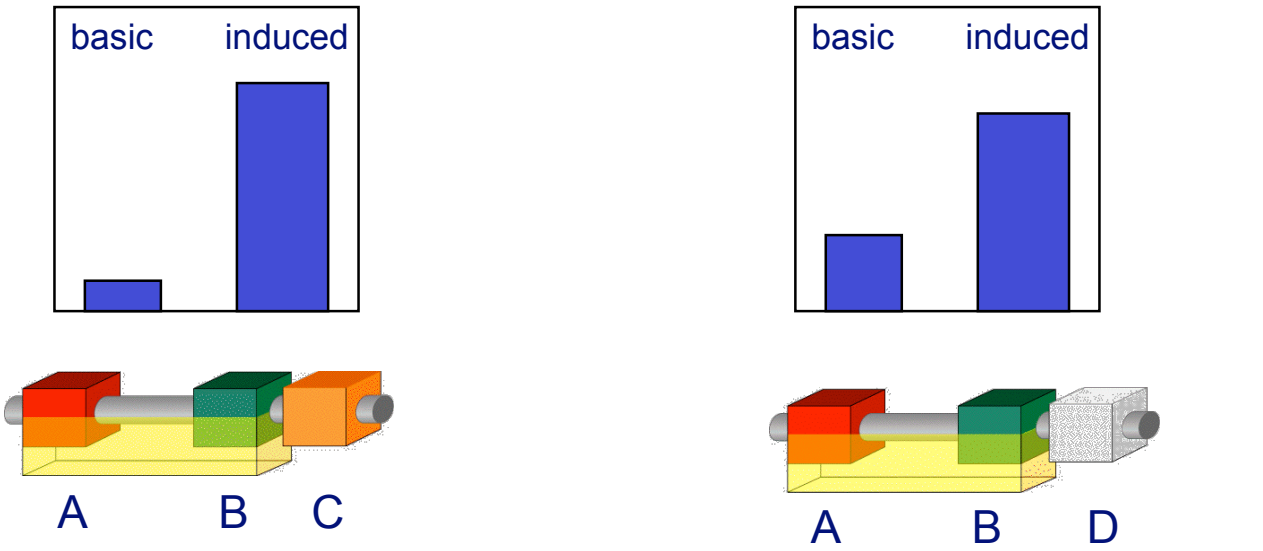

Deletions in native context do not necessarily define the complete module!


© 2004 by Genomatix Software GmbH

# Part III: Transcriptional modules

## How can a complete module be defined?

- Transfer module to a *heterologous* promoter context!




Native promoter context

Heterologous promoter context

Successful transfer of module function proves completeness of the module


© 2004 by Genomatix Software GmbH

# Part III: Transcriptional modules

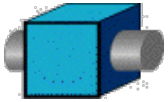
## Modules contribute strongly to functional promoter analysis

- Modules are usually linked to at least one known biological function
- A module match in a promoter makes this gene a good *candidate*
- A module match in a promoter does not prove the gene to be a target
- Additional independent evidence is required to prove the target
- A module match immediately suggests experimental verification

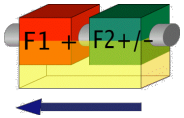
Module matches reduce experimental efforts by orders of magnitude


© 2004 by Genomatix Software GmbH

# Promoter analysis strategies



Transcription factor binding sites




Transcriptional modules

Comparative genomics of promoters

**Understanding Promoter analysis results Part III**

General strategy

Appropriate Controls

What frameworks can do and can‘t do



© 2004 by Genomatix Software GmbH

## Promoters can be phylogenetically conserved

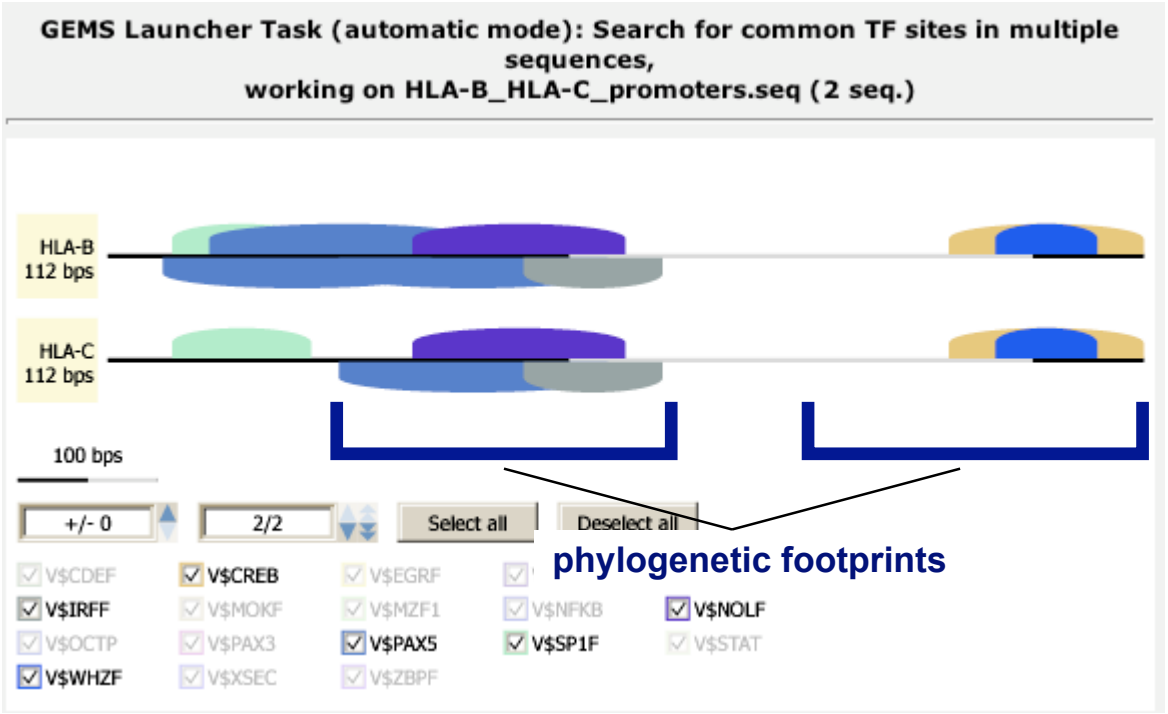
- Promoters containing *long* phylogenetic footprints are not informative
  **(Common TFs, even frameworks become a trivial consequence of sequence identity)**

- Orthologous promoters are not necessarily functionally orthologous
  **(When the gene evolved to new or distinct function, promoter structure differs)**

Orthologous promoter analysis yields basic promoter-specific frameworks


© 2004 by Genomatix Software GmbH

## Phylogenetic footprints in promoters




Frameworks are most meaningful when the sequences are not similar


© 2004 by Genomatix Software GmbH

## Orthologous promoters can be functionally different

Othologous genes do not necessarily fulfill orthologous functions

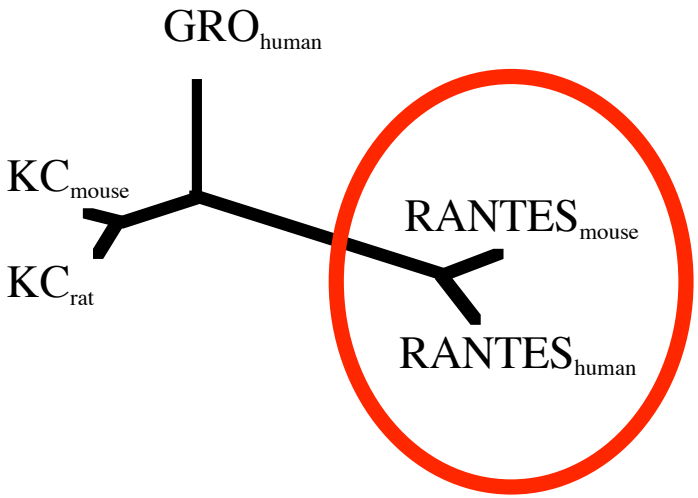
- Human and Mouse RANTES are othologous genes
- Mouse RANTES binds to a different receptor than human RANTES
- This activity is mediated by GRO in humans
- Mouse RANTES and human GRO appear to be functionally similar

Is human RANTES functionally similar to rodent KC (Gro ortholog)?

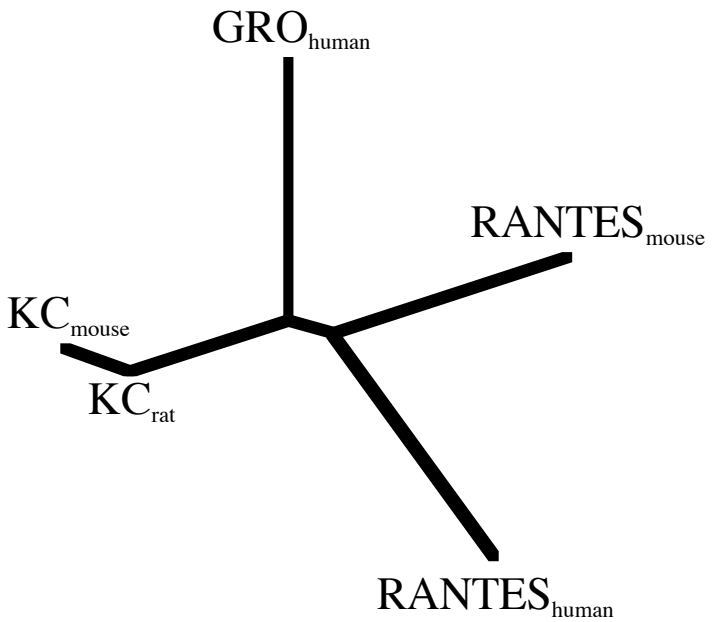

© 2004 by Genomatix Software GmbH

## Functional shift is not apparent on sequence level

Protein sequence




Promoter sequence

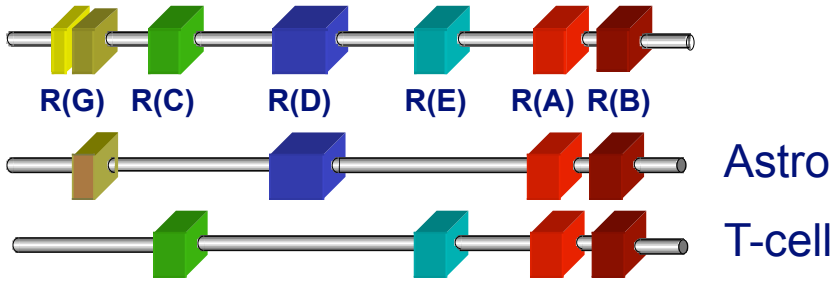



The functional similarities are not correctly reflected in sequence similarities


© 2004 by Genomatix Software GmbH

## Functional shift is apparent from promoter organization

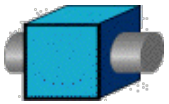
| Gene (accession #) | Models: Astro | Tcell | Mesangial | Mono | Fibro |
|---|---|---|---|---|---|
| RANTES$_{human}$ (AB023652) | ● | ● | ● | ● | ● |
| GRO/KC$_{rat}$ (U85628) | ● | ● | | | |
| GRO/KC$_{mouse}$ (S79767) | | ● | | | |
| RANTES$_{mouse}$ (U02298) | | | | | |



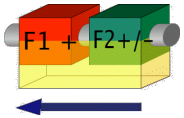


© 2004 by Genomatix Software GmbH

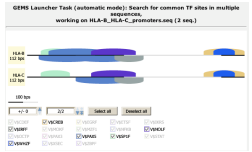
# Promoter analysis strategies



Transcription factor binding sites




Transcriptional modules



Comparative genomics of promoters

**Understanding Promoter analysis results Part III**

General strategy

Appropriate Controls

What frameworks can do and can‘t do



© 2004 by Genomatix Software GmbH

# Part III: What frameworks can do

## TFBS frameworks bear directly on functional aspects

- Frameworks are independent of direct sequence similarity
- Frameworks are molecular endpoints of signaling pathways
- Frameworks purge candidate lists by orders of magnitude
- Frameworks directly generate functional hypotheses
- Frameworks do not prove functional similarity

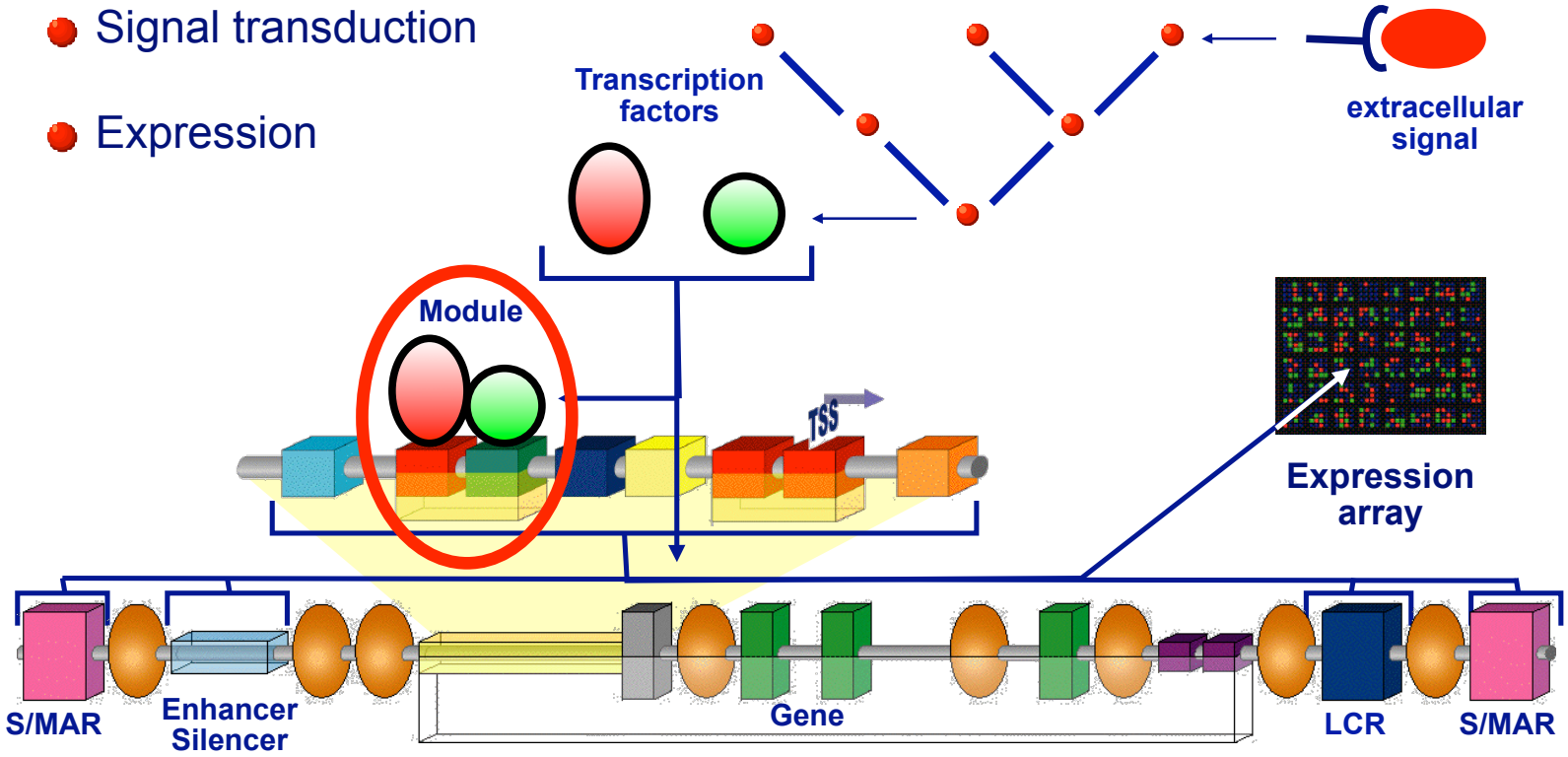
The significance and functional meaning of TFBS frameworks must be assessed


© 2004 by Genomatix Software GmbH

# Part III: What frameworks can do

## Frameworks are molecular endpoints of signaling pathways

- Signal transduction
- Expression




Promoter analysis reveals molecular mechanisms of gene expression


© 2004 by Genomatix Software GmbH

## Frameworks purge candidate lists by orders of magnitude

Genome-wide search

| Search item | total matches | % true positives* |
|---|---|---|
| Binding sites | Millions | ≤ 0.01 % |
| Frameworks (2 sites) | Thousands | ≤ 10 % |
| Frameworks (>2 sites) | 10 - hundreds | 10% - 100 % |

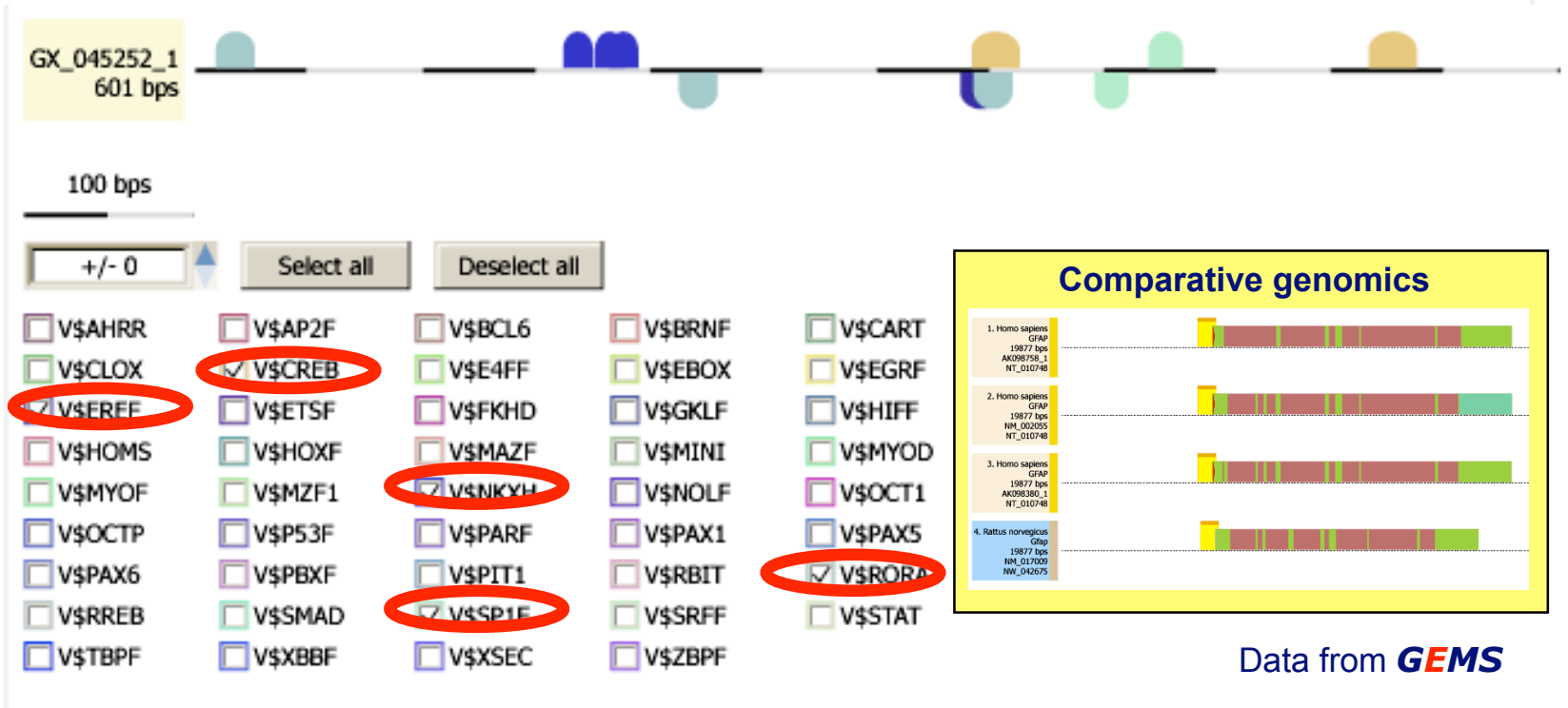
*assuming ≤ 100 true positives per factor

Frameworks are a prerequisite for elucidation of molecular networks


© 2004 by Genomatix Software GmbH

# Part III: What frameworks can do

## Frameworks directly suggest functional hypotheses




Data from **GEMS**

The framework reduces the TFBS list thus providing functional hints

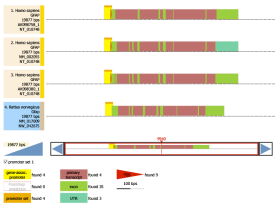


© 2004 by Genomatix Software GmbH

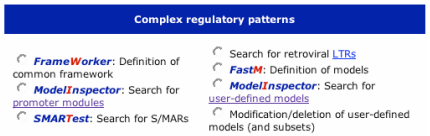
# Part III: What frameworks can do

## The Glial Fibrillary Acidic Protein (GFAP) promoter framework

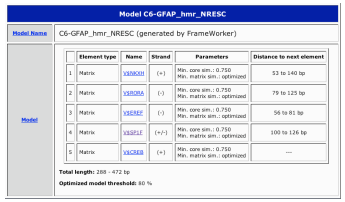
- Gold standard: promoter model derived from human, mouse and rat genes



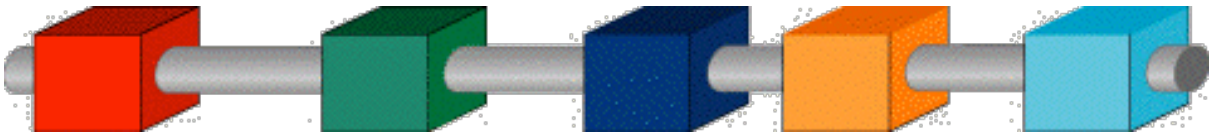
**ElDorado**
Comparative genomics




**GEMS**
**FrameWorker**




**GEMS**
**ModelInspector**




This framework is absolutely GFAP promoter specific


© 2004 by Genomatix Software GmbH

# Part III: What frameworks can do

## The Glial Fibrillary Acidic Protein (GFAP) promoter framework

- No common predefined module was found in these promoters

**Model C6-GFAP_hmr_NRESC**

| | |
|---|---|
| Model Name | C6-GFAP_hmr_NRESC (generated by FrameWorker) |

Model

| | Element type | Name | Strand | Parameters | Distance to next element |
|---|---|---|---|---|---|
| 1 | Matrix | V$NKXH | (+) | Min. core sim.: 0.750<br>Min. matrix sim.: optimized | 53 to 140 bp |
| 2 | Matrix | V$RORA | (-) | Min. core sim.: 0.750<br>Min. matrix sim.: optimized | 79 to 125 bp |
| 3 | Matrix | V$EREF | (-) | Min. core sim.: 0.750<br>Min. matrix sim.: optimized | 56 to 81 bp |
| 4 | Matrix | V$SP1F | (+/-) | Min. core sim.: 0.750<br>Min. matrix sim.: optimized | 100 to 126 bp |
| 5 | Matrix | V$CREB | (+) | Min. core sim.: 0.750<br>Min. matrix sim.: optimized | --- |

**Total length:** 288 - 472 bp

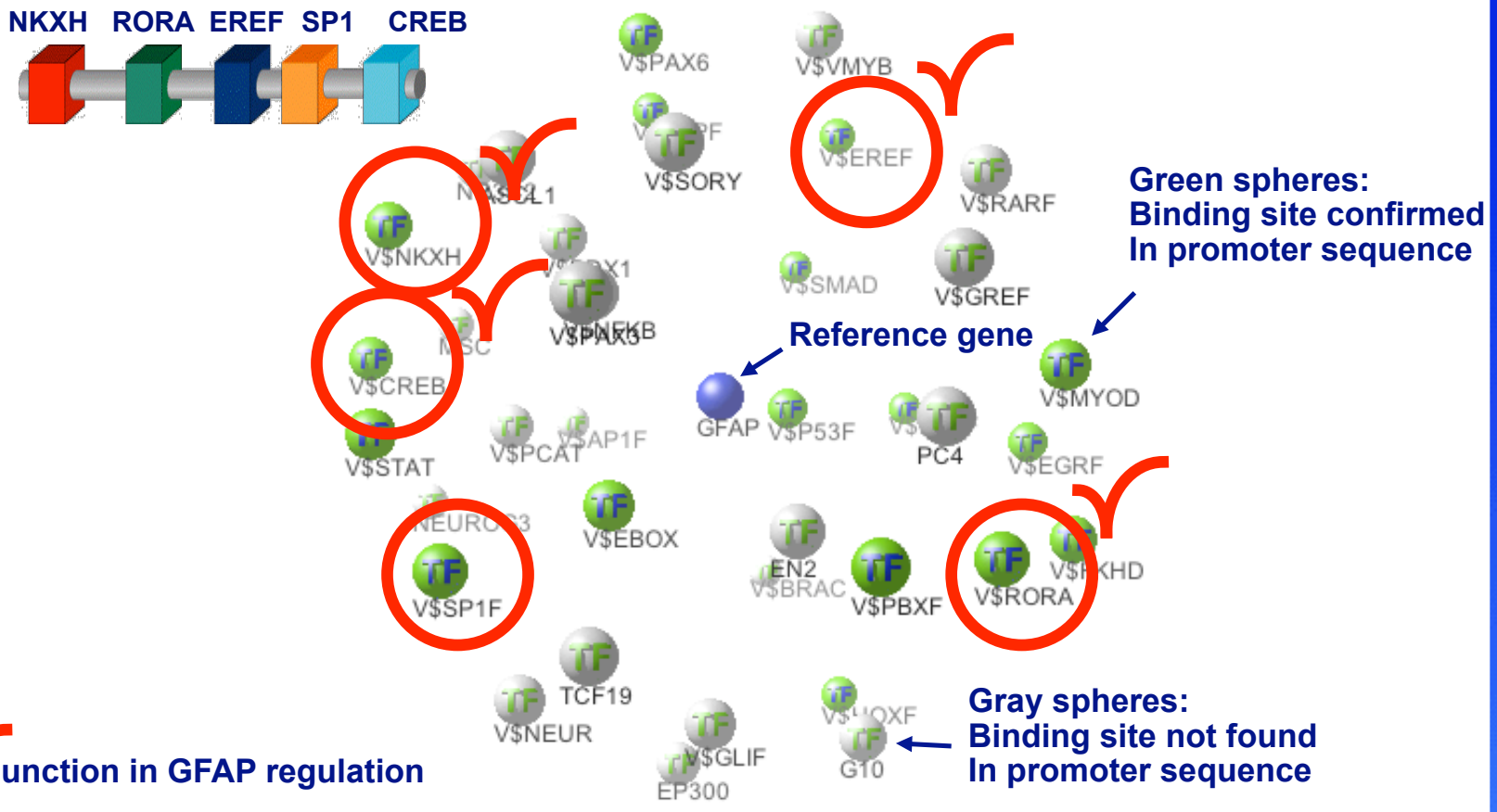
**Optimized model threshold:** 80 %

**These sites were found in module matches (only in pairs of sequences)**



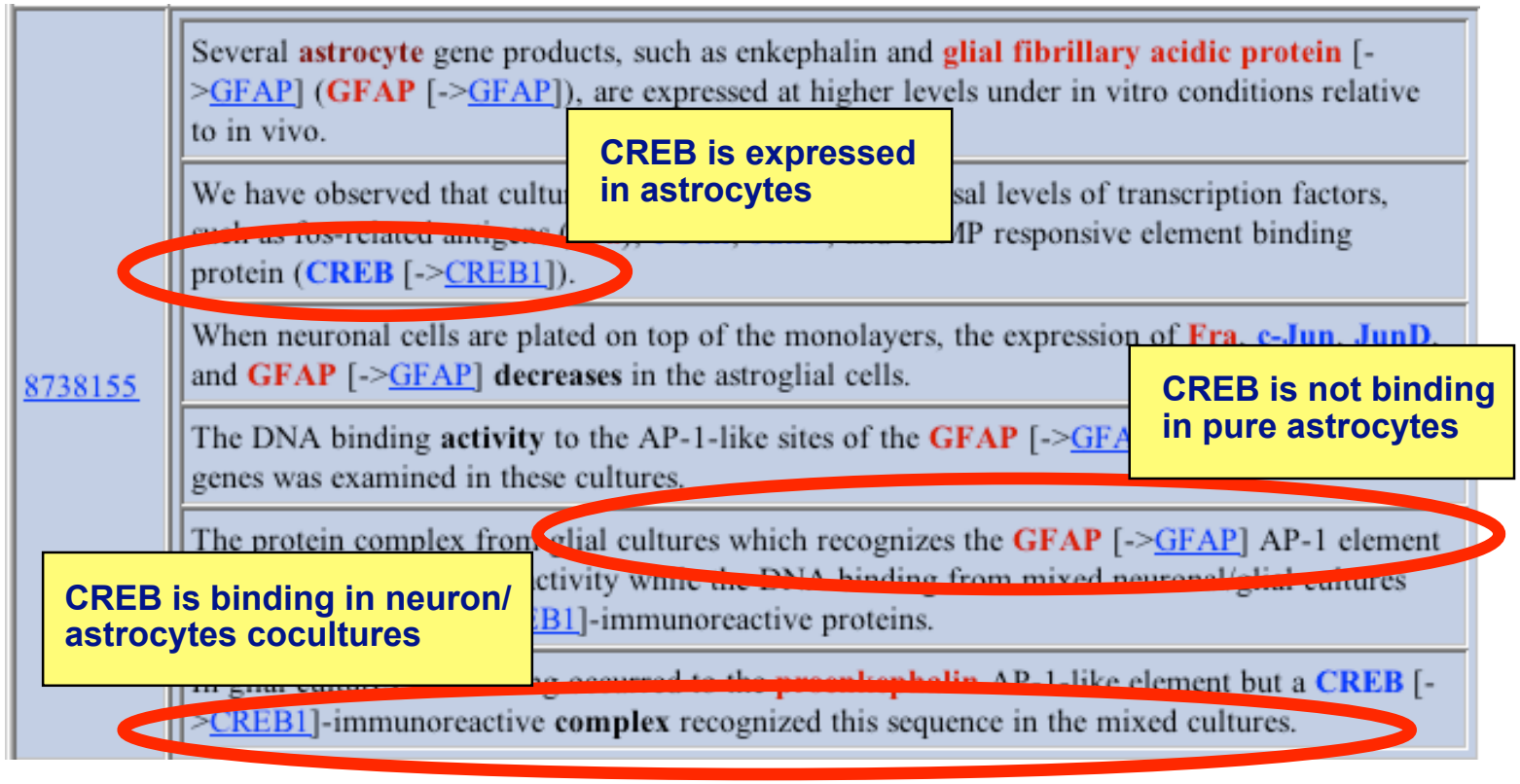
© 2004 by Genomatix Software GmbH

## *BiblioSphere* analysis of the GFAP gene TF relations




© 2004 by Genomatix Software GmbH

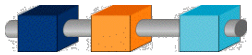
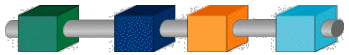
## *BiblioSphere* GFAP - CREB cocitation




© 2004 by Genomatix Software GmbH

# Part III: What frameworks can do

## The Glial Fibrillary Acidic Protein (GFAP) promoter framework

| Model used in search<br>**Threshold** | Matches in EPD<br>**(≈ 3,000 promoters)** | Matches in Human promoters<br>**(≈ 59,000 promoters)** |
|---|---|---|
| **Default**  | 417 | 5325 |
|  | 1 | 22 |
|  | 0 | 3 |
|  | 0 | 3 |
| **Relaxed**  | 1 | 14 |

Shortening of framework or relaxing thresholds yields different results



© 2004 by Genomatix Software GmbH

Genomatix

# Glial fibrillary acid protein

## The Glial Fibrillary Acidic Protein (GFAP) promoter framework

- 14 matches in human promoters
- 3 matches were GFAP transcripts
- 6 matches were unannotated transcripts
- 5 matches were annotated transcripts:

| | |
|---|---|
| - RDHL **(NADP-dependent retinol dehydrogenase)** | No annotation available |
| - ENNP4 **(ectonucleotide pyrophosphatase)** | **Coexpressed with GFAP in astrocytes** |
| - PRDX4 **(perioxireduxin)** | No connection detcted |
| - NDN **(necdin homolog)** | **Binds toE2F which regulates GFAP** |
| - DGAT2 **(diacylglycerol O-acetyltransferase homolog)** | **DGAT2 &GFAP are insulin induced** |

Data from ***BiblioSphere***

Three of five genes are indirectly linked to GFAP

© 2004 by Genomatix Software GmbH

# Part III: What frameworks can do

## Summary

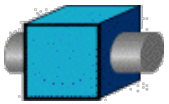
TFBS frameworks help to trace transcription control genome-wide

- TFBS frameworks & modules are the endpoints of signaling pathways
- TFBS frameworks reduce the search space for experimental design
- TFBS frameworks can be unique fingerprints of promoters/enhancers
- One promoter region can contain several distinct TFBS frameworks

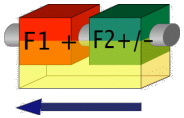
A framework resembles a description of one or more promoter functions


© 2004 by Genomatix Software GmbH

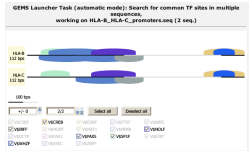
# Promoter analysis strategies



Transcription factor binding sites




Transcriptional modules

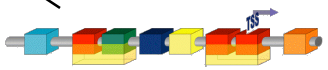


Comparative genomics of promoters

**Understanding Promoter analysis results Part III**

General strategy

Appropriate Controls




What frameworks can do and can‘t do



© 2004 by Genomatix Software GmbH

# Part III: Appropriate controls

## Computational results can be misleading

Computational results need to be as carefully controlled as lab experiments

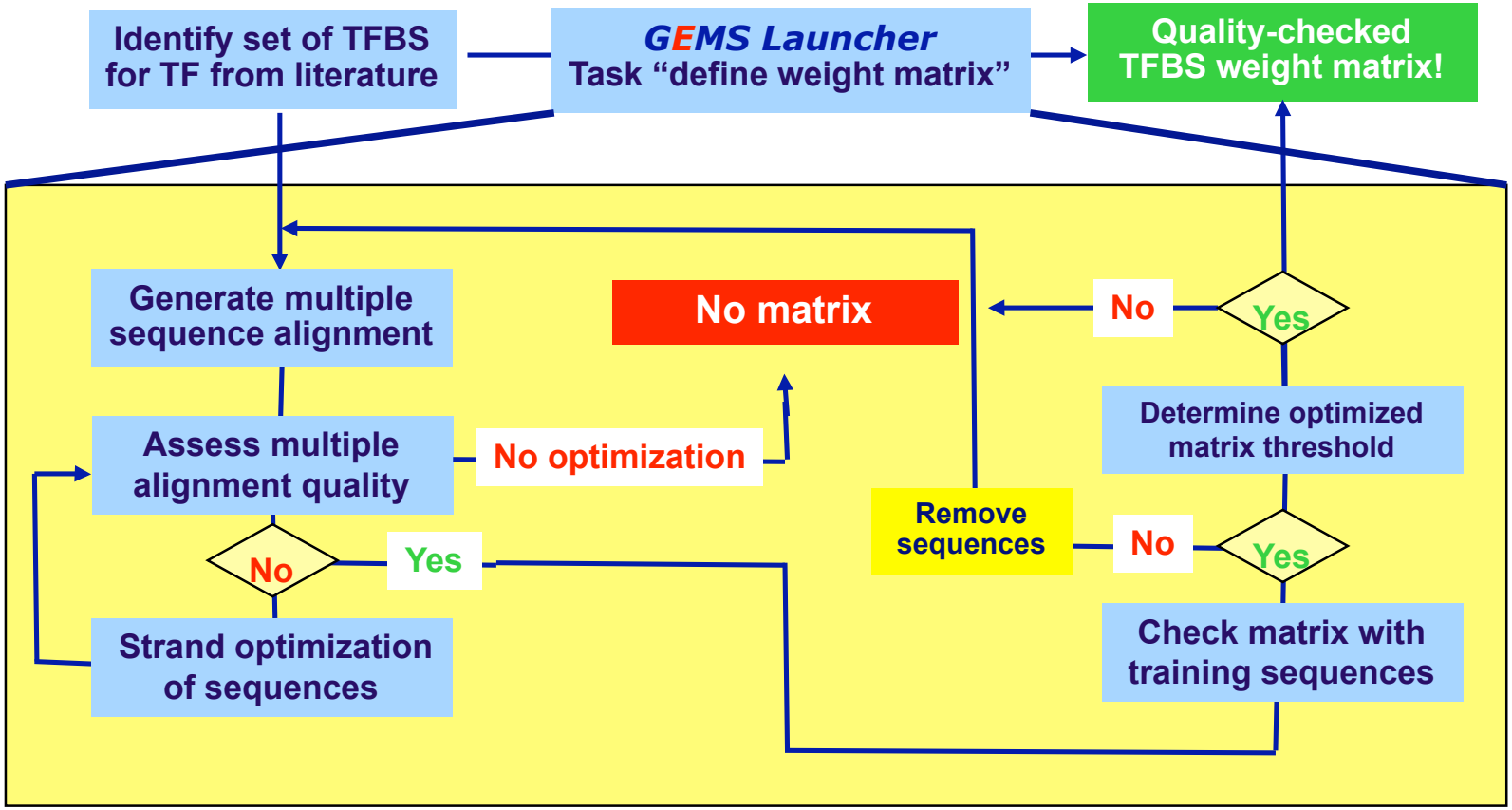
- Never rely on results from only one type of analysis
  **(Always seek confirmation of results from a different source)**
- The right choice of tools is critical
  **(same as deciding on a suitable experimental setup)**
- Appropriate controls have to be applied to all results
  **(same as checking ingredients individually before experiments)**
- Biological relevance meets statistical significance in the *right* context
  **(a weak binding site may be spurious, in the context of a module it becomes crucial)**
- Several whole genomes are available - use them!
  **("sheltered" environment of a single or a few hand-picked examples is not acceptable)**

Inappropriate controls provide false safety and are worse than no controls!


© 2004 by Genomatix Software GmbH

# Part III: Appropriate controls

## Generate a weight matrix for a TFBS






© 2004 by Genomatix Software GmbH

# Part III: Appropriate controls

## Biologically meaningful control sequences are crucial

- Functionally verified sequences (true or false) are scarce
- Random DNA can be easily generated in quantity
  - Random DNA fails to reflect important biological features
  - Underrepresented, repetitive or symmetric elements
- Large genomic DNA provides an “all-purpose” control
  - Natural mixture of real elements
  - Provides competitive elements in natural proportions

The human genome contains about 3 billion bps, but almost no random DNA


© 2004 by Genomatix Software GmbH

# Appropriate controls

**Unsuitable control sequences can bias results considerably!**

- Example 1: Specificity of TF-binding site detection
  - Control set: bacterial DNA (for mammalian TF-sites)
  - “High” specificity due to different nucleotide composition

=> Other mammalian TF binding sites and mammalian genomic sequences!

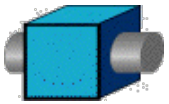
- Example 2: Promoter recognition
  - Control set: coding sequences
  - “Low” false positives due to codon restrictions

=> Unrelated promoters or non-coding sequences from same species!

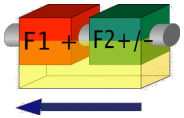
Control sequences should be as similar to the “true” set as possible!


© 2004 by Genomatix Software GmbH

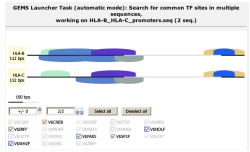
# Promoter analysis strategies



Transcription factor binding sites




Transcriptional modules

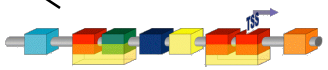


Comparative genomics of promoters

Understanding Promoter analysis results Part III

General strategy

Appropriate Controls




What frameworks can do and can‘t do



© 2004 by Genomatix Software GmbH

# Part III: General strategy

## A quest for targets of pathways acting through Pax 8

Facts

- Pax 8 is a transcription factor binding directly to promoters
- Pax 8 binding sites are available from known target genes

Questions

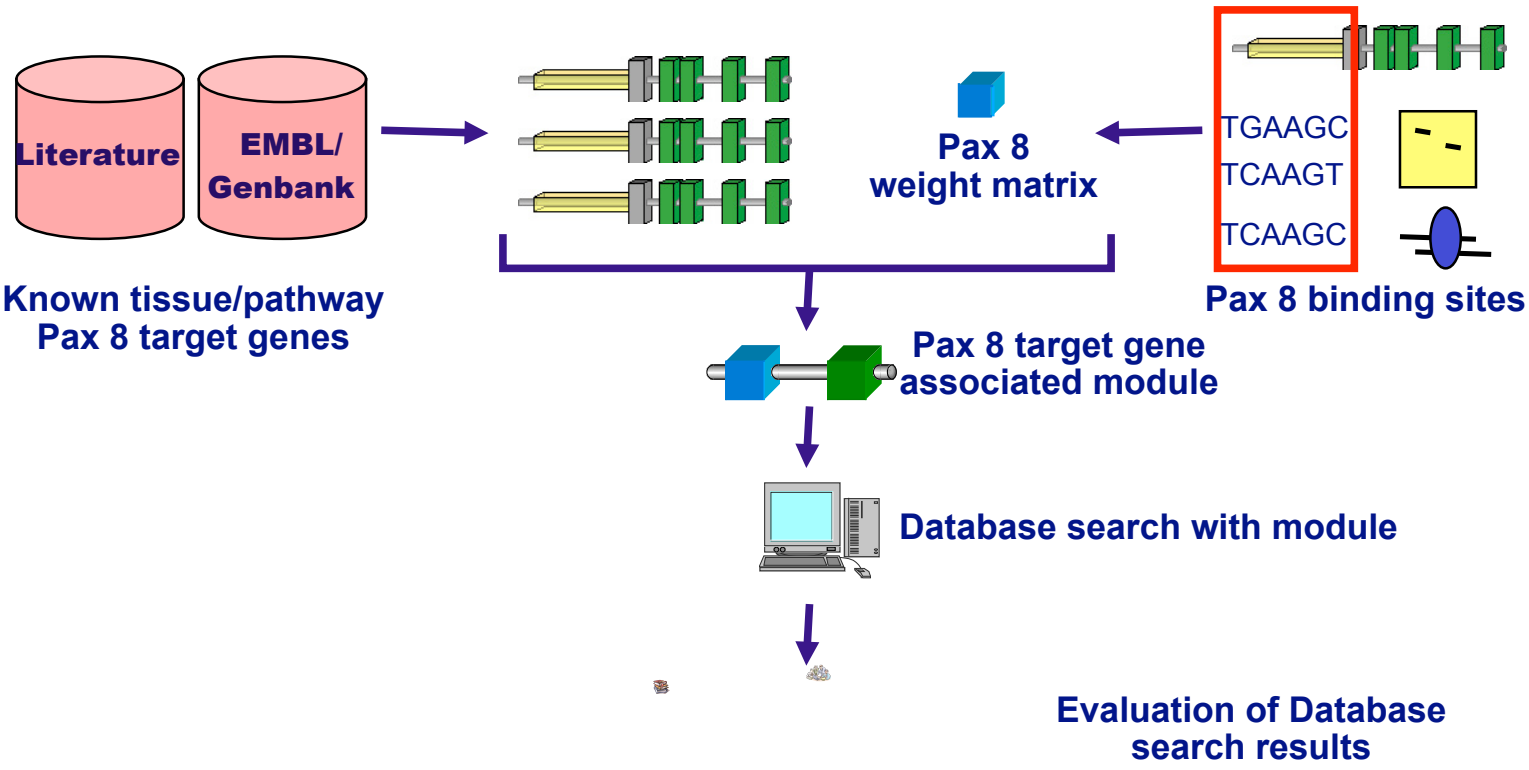
- How to find Pax 8 binding sites in new promoters?
- Since Pax 8 has no “target” genes, what else defines a target gene?
- How to find potential new target genes?

Three questions - one answer: Promoter analysis!


© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Strategy for finding Pax 8 "target" genes



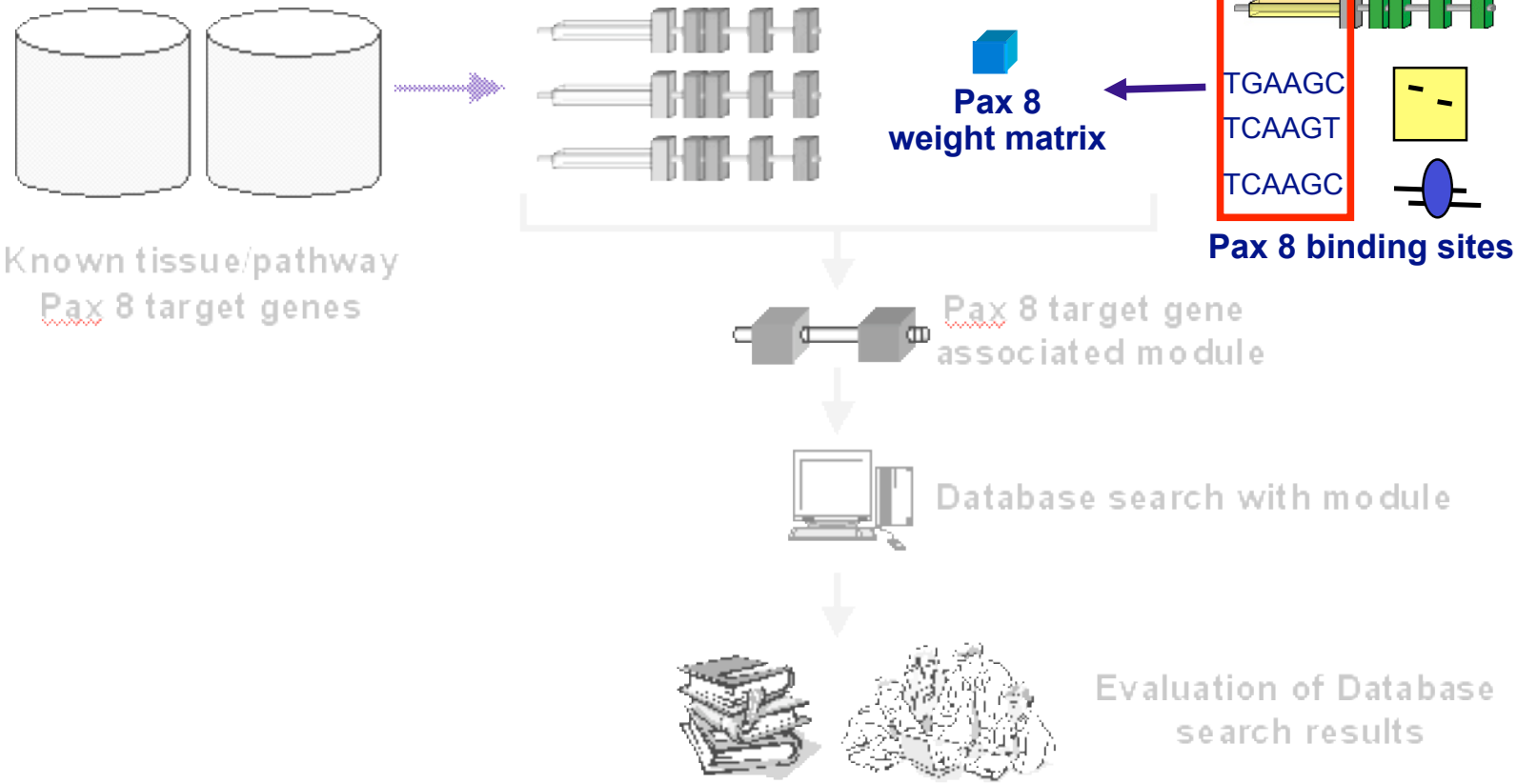

This strategy will yield targets for a *single* pathway involving Pax 8



© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Step 1




Generating a weight matrix for Pax 8

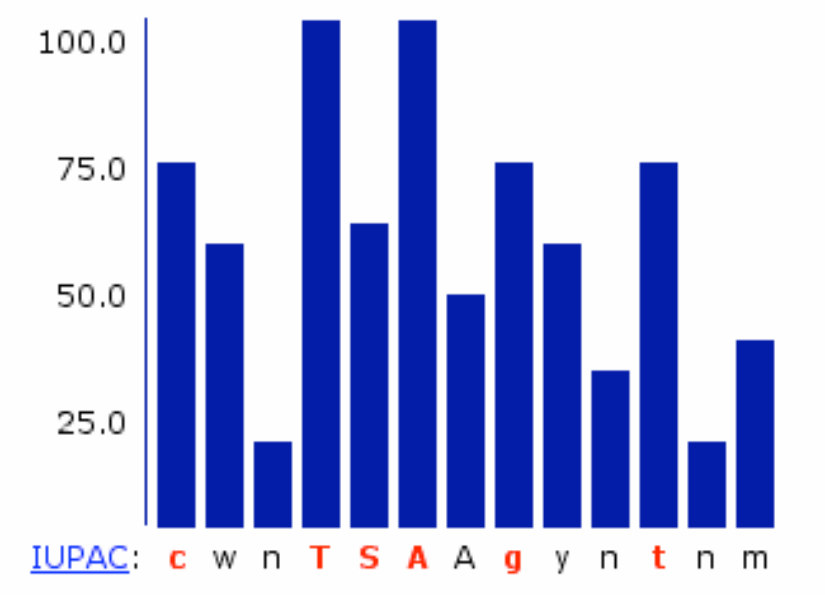


© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Generation of a Pax 8 binding site weight matrix

- 8 known Pax 8 binding sites extracted from literature
- ***GEMS*** generated a specific weight matrix





© 2004 by Genomatix Software GmbH

# Part III: General strategy

## The Pax 8 binding site is not sufficient

**GEMS**

| | |
|---|---|
| **ElDorado** promoters (human and mouse)<br>Pax 8 matrix | 31,802 |

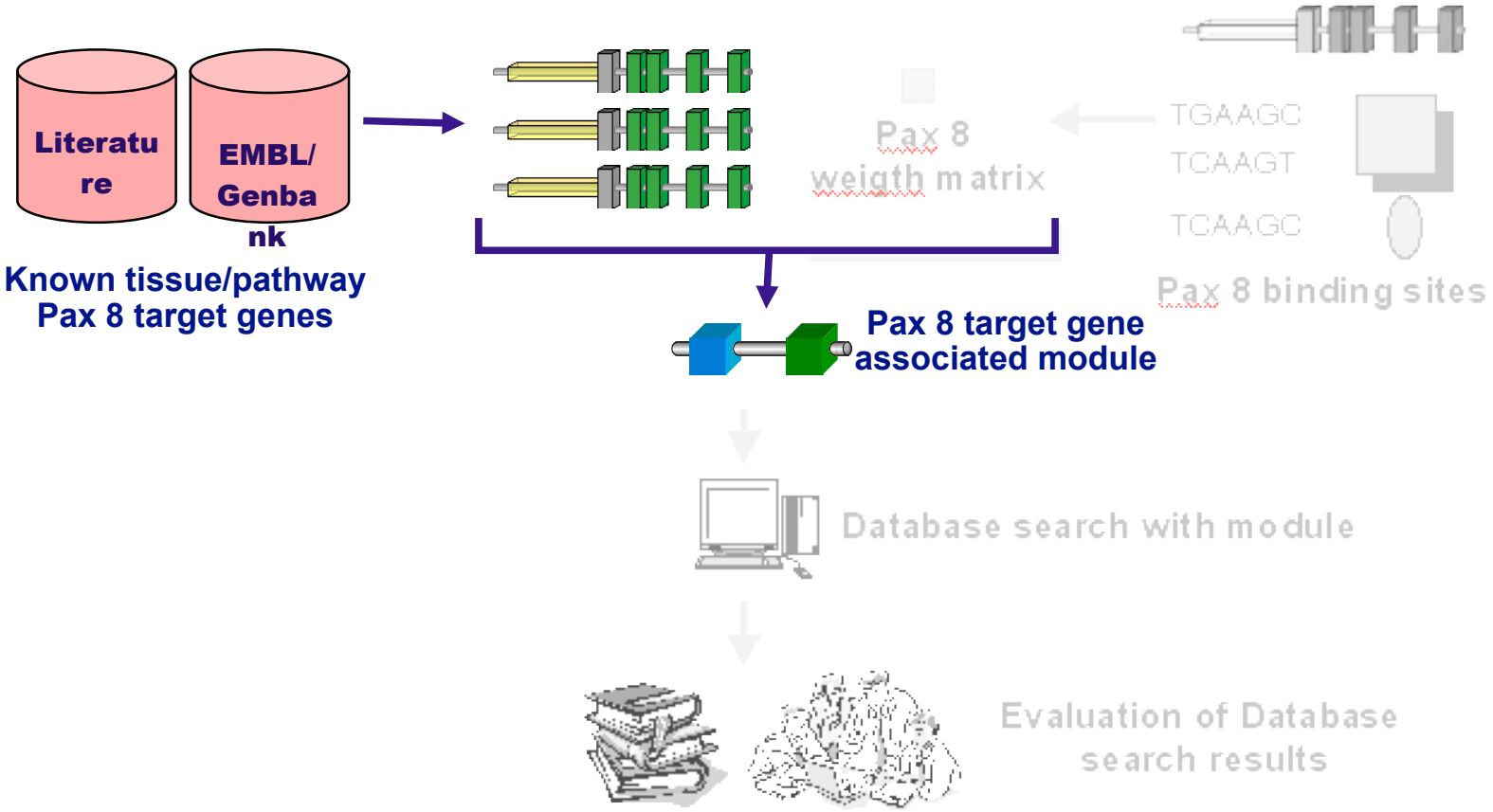
- Pax 8 binding sites were located with optimized thresholds
- Most of these sites are actually likely to bind the Pax 8 protein

The Pax 8 binding site alone is insufficient to predict target genes!


© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Step 2 & 3




Analysis of known target genes involving Pax 8, definition of target module



© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Pax 8 is reported to be involved in thyroid specific expression

- Literature analysis identifies another transcription factor to be involved
- This other factor is the thyroid transcription factor TTF1
- TTF1 and Pax 8 act synergistically in at least two promoters

✗ Problem: There was not TTF1 weight matrix available

- 8 known TTF-1 binding sites from literature
- Weight matrix defined by ***GEMS***

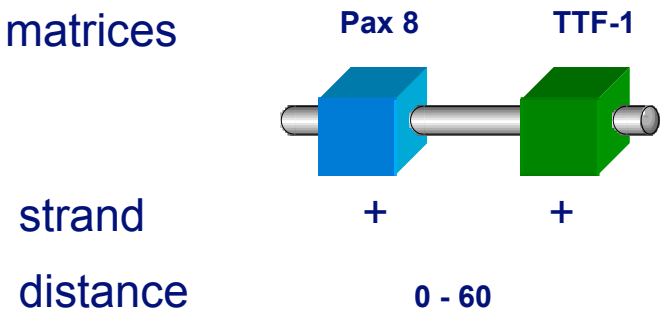
Pax 8 and TTF1 are (part of) a promoter module defining target genes


© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Some thyroid specific Pax 8 target contain a promoter module

- Thyroglobulin and thyroperoxidase promoters share a module



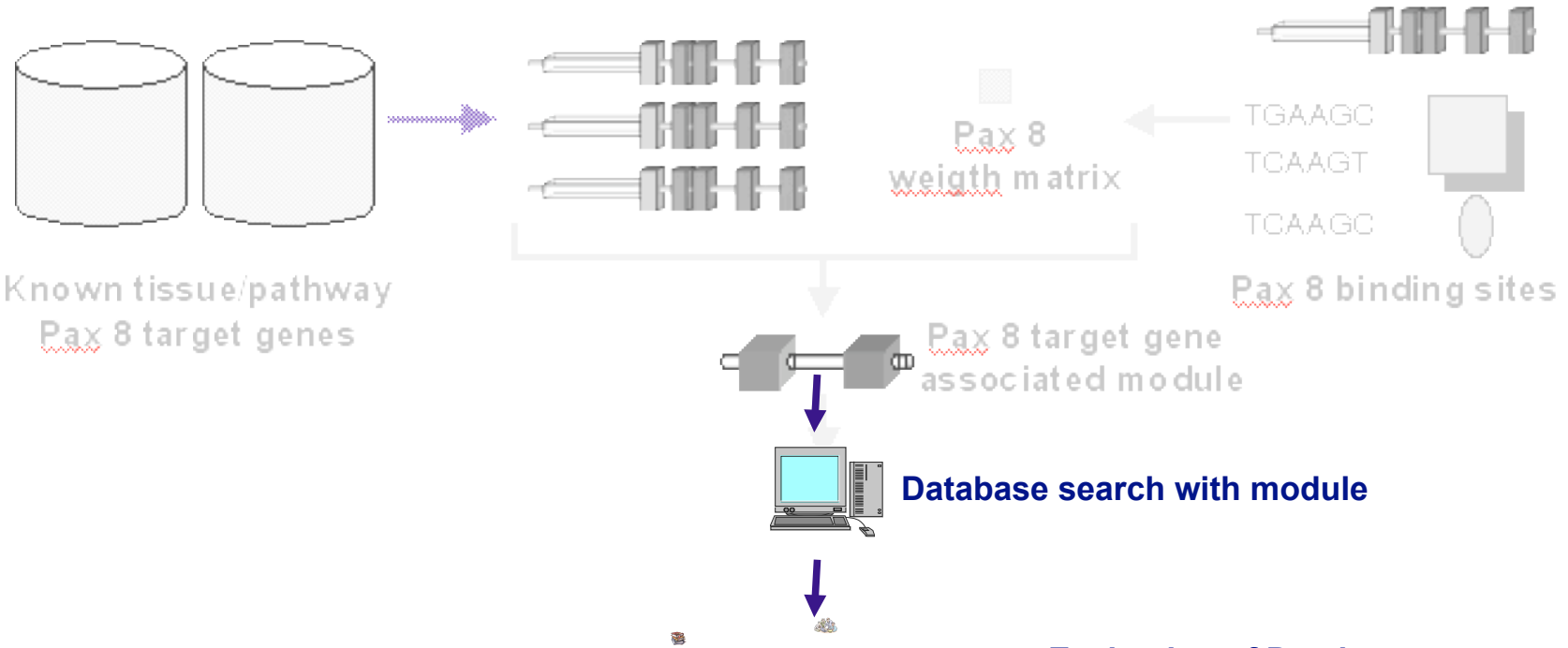

- This module recognizes functional sites in known target promoters

Zannini 1992, Mol Cell Biol 12, 4230.


© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Step 3 & 4




Database search & evaluation of results



© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Binding site matrix versus promoter module

***GEMS***

***ElDorado*** promoters (human and mouse)

| | |
|---|---|
| Pax 8 matrix | 31,802 |
| TTF-1 matrix | 19,829 |
| Pax 8 / TTF-1 module | 418 |

| | |
|---|---|
| Pax 8 matrix | 1 match per ~ 2,000 |
| TTF-1 matrix | 1 match per ~ 3,000 |
| Pax 8 / TTF-1 module | 1 match per ~ 144,000 |

The module is almost 2 orders of magnitude more selective than the matrix



© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Reducing the match list by annotation filtering

- Extraction of biologically relevant targets by keyword search (thyroid)

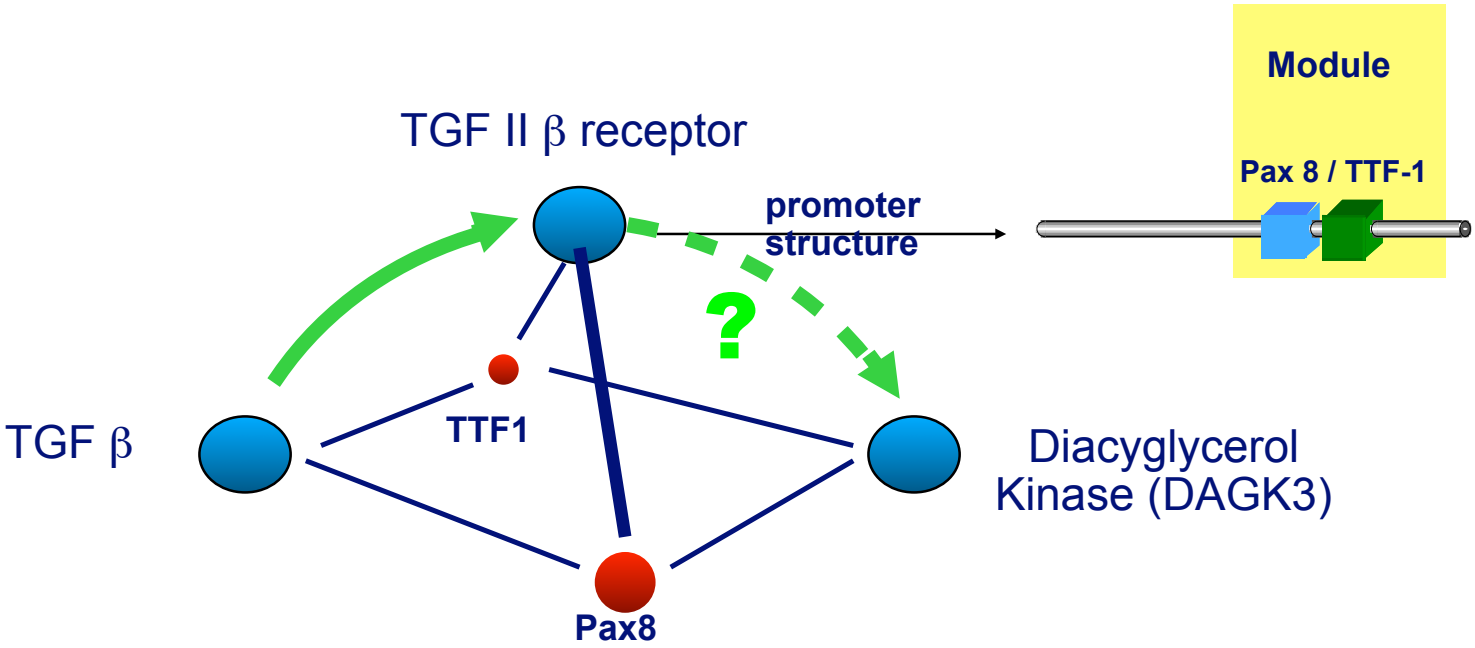
| Sequence | Model Name | Position | Strand | Model Score |
|---|---|---|---|---|
| **AF085346** [AF085346] (1 - 2473)<br>**[DNA] Homo sapiens involucrin gene, upstream regulatory region.** | PAX8_TTF1 | 1530 - 1547 | (+) | 90.4 % |
| **AF186257** [AF186257] (1 - 501)<br>**[DNA] Homo sapiens sulfotransferase 1C1 (SULT1C1) gene, exon 1.** | PAX8_TTF1 | 369 - 352 | (-) | 93.2 % |
| **HS17170** [U17170] (1 - 954)<br>**[DNA] Human type II transforming growth factor-beta receptor gene, promoter region.** | PAX8_TTF1 | 488 - 424 | (-) | 91.0 % |
| **HSAMD01** [M88003] (1 - 3159)<br>**[DNA] Human S-adenosylmethionine decarboxylase (AMD1) gene, exon 1.** | PAX8_TTF1 | 514 - 531 | (+) | 92.4 % |
| **HSMBP1A1** [M63599] (1 - 2593)<br>**[DNA] Human myelin basic protein (MBP) gene, exon 1.** | PAX8_TTF1 | 1621 - 1604 | (-) | 90.3 % |



© 2004 by Genomatix Software GmbH

# Part III: General strategy

## TGF II β receptor is a potential target gene

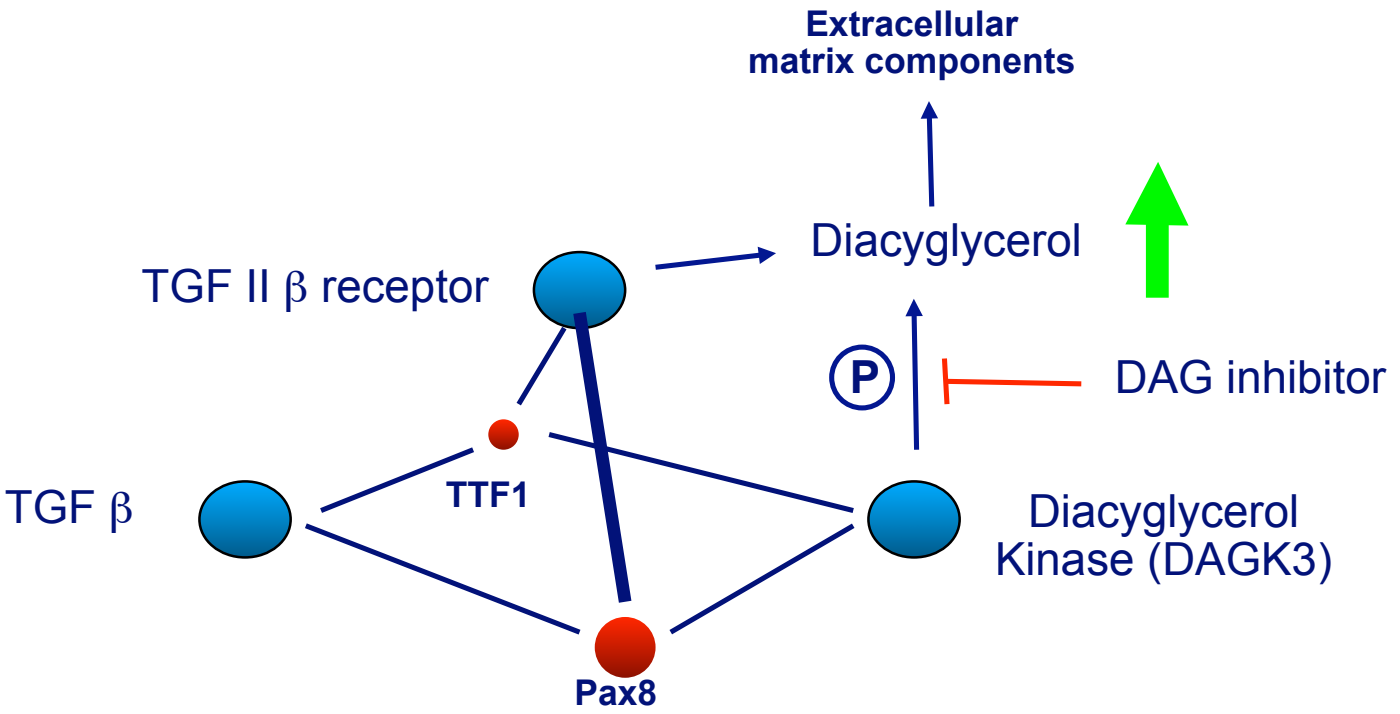



Transcriptional modules are crucial elements of regulatory networks


© 2004 by Genomatix Software GmbH

# Part III: General strategy

## TGF II β receptor action may be regulated by a common module




Transcriptional modules can organize regulatory feedback loops



Ignotz & Honeyman (2000), J Cell Biochem. 7, 588-594.

© 2004 by Genomatix Software GmbH

# Part III: General strategy

## Summary

The type of input data determines the questions that can be asked

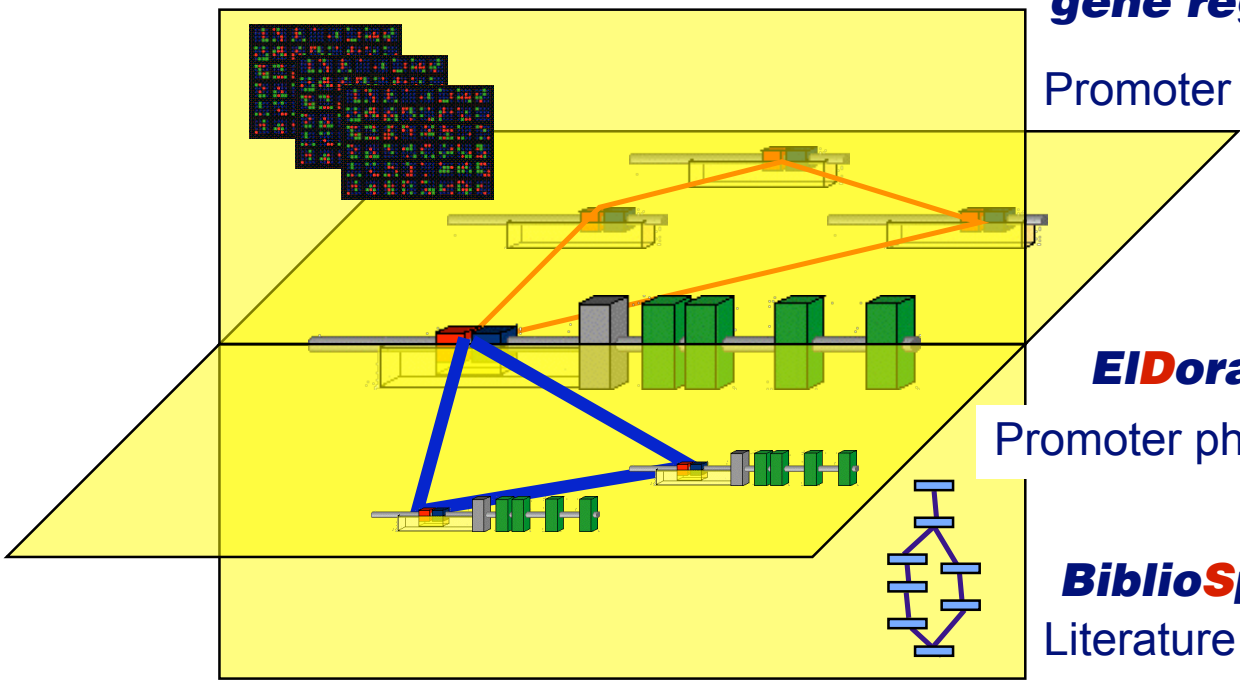
- TFBS sequences can only yield TFBS descriptions - not functions
- A set of real target genes can yield a pathway target *module*
- Such *modules* or frameworks can be used to locate new candidates
- Often networks or feedback loops can be identified by *modules*

Regulatory networks link genes functionally independent of physical interaction


© 2004 by Genomatix Software GmbH

# Part II: Regulatory networks

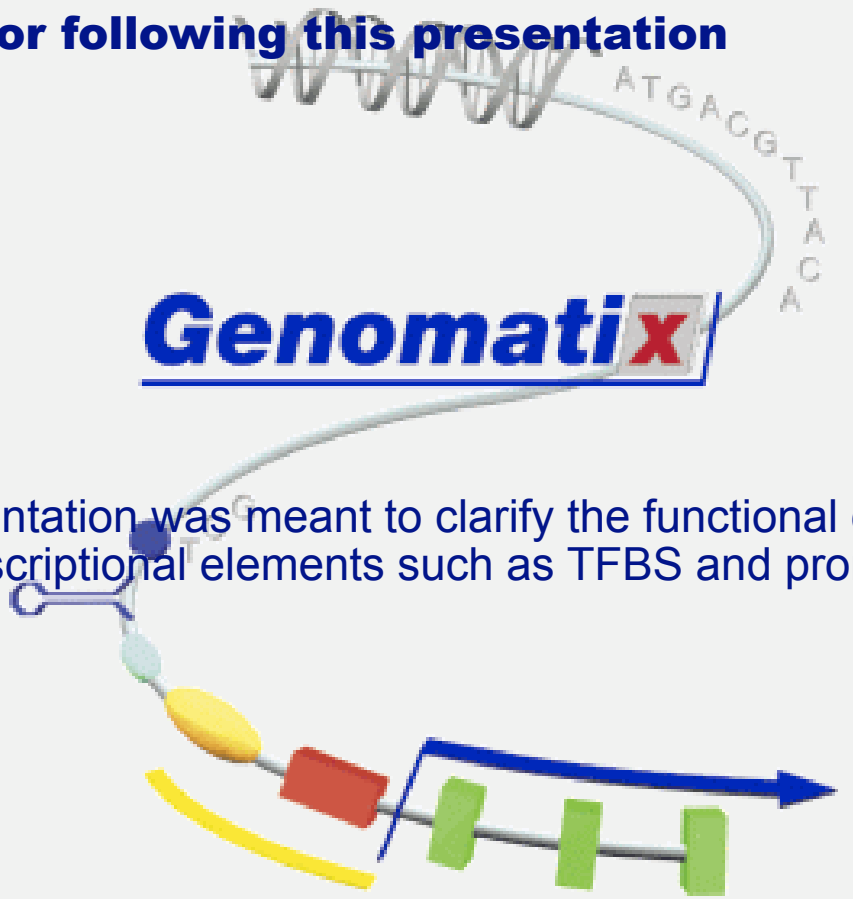
## Independent data enhance functional context analysis




Only independent data and algorithms allow a synergistic consensus approach


© 2004 by Genomatix Software GmbH

## Thank you for following this presentation




This presentation was meant to clarify the functional complexity of transcriptional elements such as TFBS and promoters.

Development of all Genomatix products is carried out according to these rules


© 2004 by Genomatix Software GmbH