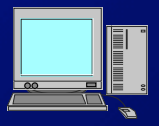
Genomatix

*Interactive workshop*



# Hands on session: Advanced promoter analysis

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

Genomatix

# Outline of tasks

## What you will be doing during the next 2 hours

- Analyze rbcS promoters for transcription factor binding sites (TFBS) (***MatInspector*** from ***GEMS*** package)
- Analyze rbcS promoters for TFBS frameworks (***FrameWorker*** from ***GEMS*** package)
- Modify and optimize rbcS promoter models for TFBS frameworks (***FastM*** from ***GEMS*** package)
- Database search with frameworks (***ModelInspector*** from ***GEMS*** package)
- Determine and define new TFBS matrix from multiple alignment (***DiAlign*** and ***MatDefine*** from ***GEMS*** package)
- Experimental design: specifically deleting a TFBS (***SequenceShaper*** from ***GEMS*** package)

# Task 1

## What you will be doing next

- Analyze rbcS promoters for transcription factor binding sites (TFBS) (***MatInspector*** from ***GEMS*** package)
- Analyze rbcS promoters for TFBS frameworks (***FrameWorker*** from ***GEMS*** package)
- Modify and optimize rbcS promoter models for TFBS frameworks (***FastM*** from ***GEMS*** package)
- Database search with frameworks (***ModelInspector*** from ***GEMS*** package)
- Determine and define new TFBS matrix from multiple alignment (***DiAlign*** and ***MatDefine*** from ***GEMS*** package)
- Experimental design: specifically deleting a TFBS (***SequenceShaper*** from ***GEMS*** package)

Genomatix

# Plant promoter modeling

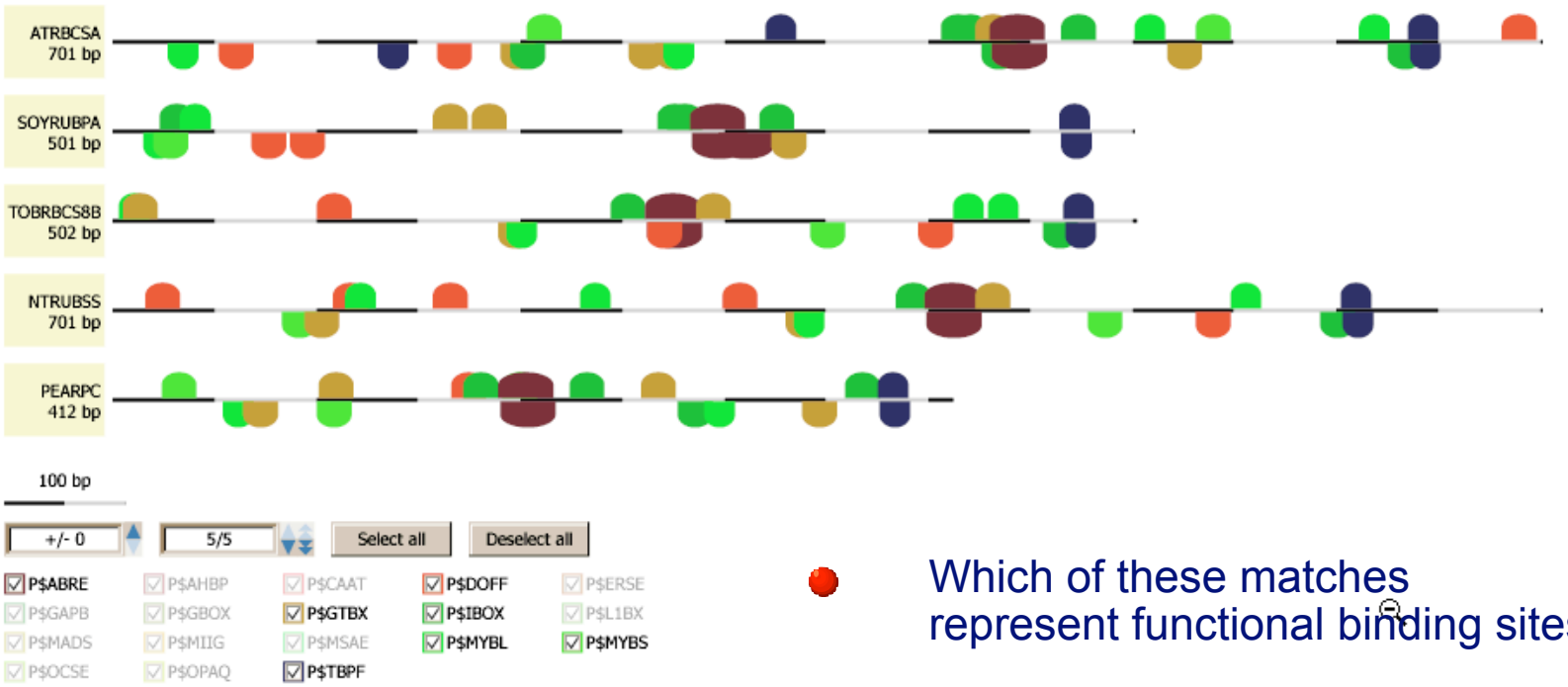
## Training dataset: five orthologous rbcS promoter sequences

- Comparative promoter analysis

| No. | Sequence Name | Sequence Description | Length |
|---|---|---|---|
| 1 | ATRBCSA | Arabidopsis thaliana ats1A gene for ribulose 1.5-biphoshate carboxylase small subunit (EC 4.1.1.39). | 701 bp |
| 2 | SOYRUBPA | Soybean ribulose 1,5-bisphosphate carboxylase small subunit (SRS4) gene, complete cds. | 501 bp |
| 3 | TOBRBCS8B | N.plumbaginifolia ribulose bisphosphate carboxylase (rbcS-8B) gene, complete cds. | 502 bp |
| 4 | NTRUBSS | Tobacco gene for ribulose 1,5-bisphosphate carboxylase small subunit. | 701 bp |
| 5 | PEARPC | Pea (P.sativum) small subunit ribulose bisphosphate carboxylase (rbcS-3A) gene, promoter region. | 412 bp |

# Search for common TF sites

**GEMS Launcher Task: Search for common TF sites in multiple sequences, working on rbcS_promoters.embl (5 seq.)**




Which of these matches represent functional binding sites?

Genomatix

# Task 2

## What you will be doing next

- Analyze rbcS promoters for transcription factor binding sites (TFBS) (***MatInspector*** from ***GEMS*** package)
- Analyze rbcS promoters for TFBS frameworks (***FrameWorker*** from ***GEMS*** package)
- Modify and optimize rbcS promoter models for TFBS frameworks (***FastM*** from ***GEMS*** package)
- Database search with frameworks (***ModelInspector*** from ***GEMS*** package)
- Determine and define new TFBS matrix from multiple alignment (***DiAlign*** and ***MatDefine*** from ***GEMS*** package)
- Experimental design: specifically deleting a TFBS (***SequenceShaper*** from ***GEMS*** package)

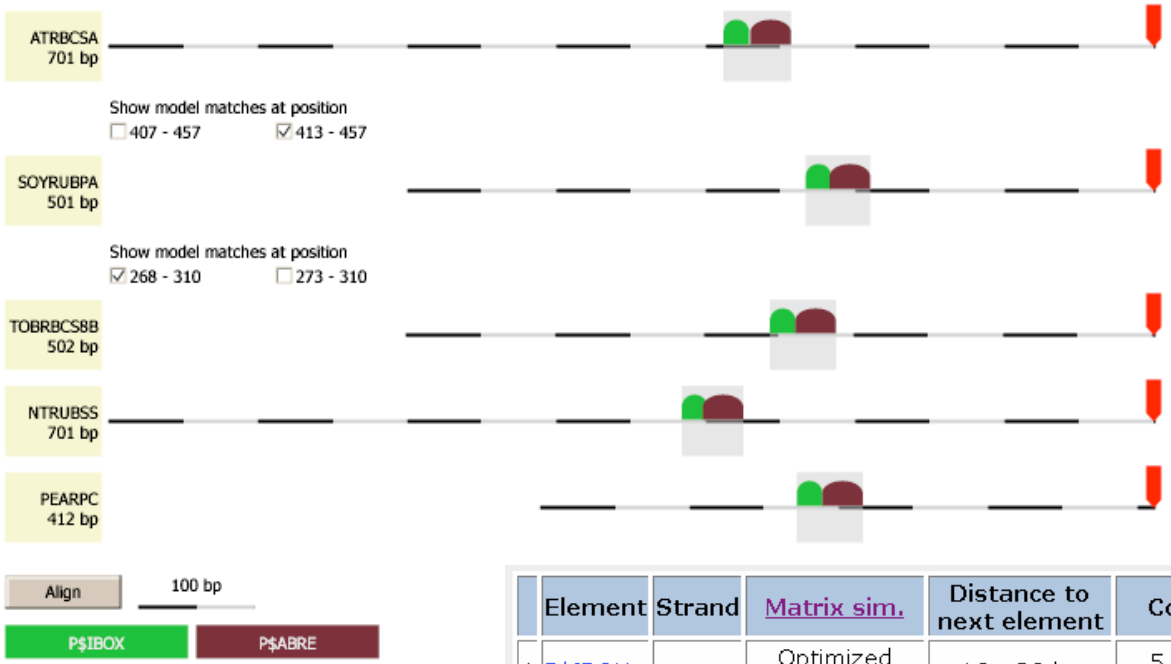
# Definition of common framework

## GEMS Launcher Task: *FrameWorker*: Definition of common framework

**FrameWorker Parameters:**

**Minimum no. of sequences with framework:** 5 sequences (100% of 5)
**Distance between elements:** 10 - 50

**Result: 1 framework with 2 elements was found:**




| | Element | Strand | Matrix sim. | Distance to next element | Common to | FW-Scores |
|---|---|---|---|---|---|---|
| 1 | P$IBOX | + | Optimized (min. 0.95) | 16 - 29 bp | 5 sequences (100 %) 7 matches, 5 non-overlapping | 0.71 / 1.00 |
| 2 | P$ABRE | + | Optimized (min. 0.73) | --- | | |

Genomatix

# Task 3

## What you will be doing next

- Analyze rbcS promoters for transcription factor binding sites (TFBS) (***MatInspector*** from ***GEMS*** package)
- Analyze rbcS promoters for TFBS frameworks (***FrameWorker*** from ***GEMS*** package)
- Modify and optimize rbcS promoter models for TFBS frameworks (***FastM*** from ***GEMS*** package)
- Database search with frameworks (***ModelInspector*** from ***GEMS*** package)
- Determine and define new TFBS matrix from multiple alignment (***DiAlign*** and ***MatDefine*** from ***GEMS*** package)
- Experimental design: specifically deleting a TFBS (***SequenceShaper*** from ***GEMS*** package)

# Modification of rbcS Model1

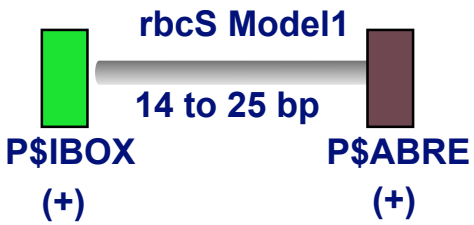
## GEMS Launcher Task: Modification of user-defined models

- Manual modification of rbcS Model1 based on matches to the five rbcS promoter training sequences (optional)

**Your model "rbcS_Model1":**

**Matrix P$IBOX  Matrix P$ABRE**

| | Name | Strand | Quality |
|---|---|---|---|
| Change element 1 | Family: P$IBOX | + | core sim. > 0.75 Optimized matrix sim. |
| Change distance range 1 | 14 to 25 bp | | |
| Change element 2 | Family: P$ABRE | + | core sim. > 0.75 Optimized matrix sim. |



Genomatix

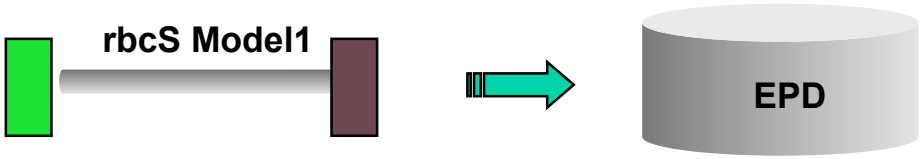
# Task 4

## What you will be doing next

- Analyze rbcS promoters for transcription factor binding sites (TFBS) (***MatInspector*** from ***GEMS*** package)
- Analyze rbcS promoters for TFBS frameworks (***FrameWorker*** from ***GEMS*** package)
- Modify and optimize rbcS promoter models for TFBS frameworks (***FastM*** from ***GEMS*** package)
- Database search with frameworks (***ModelInspector*** from ***GEMS*** package)
- Determine and define new TFBS matrix from multiple alignment (***DiAlign*** and ***MatDefine*** from ***GEMS*** package)
- Experimental design: specifically deleting a TFBS (***SequenceShaper*** from ***GEMS*** package)

# Evaluation of rbcS Model1

**GEMS Launcher Task: *ModelInspector*: Search for user-defined models, working on Buchers EPD (Eukaryotic Promoter Database, Rel. 78)**




***ModelInspector* Result: A total of 37 matches was found. Sequences searched: 4810 (2,886,000 bp).**

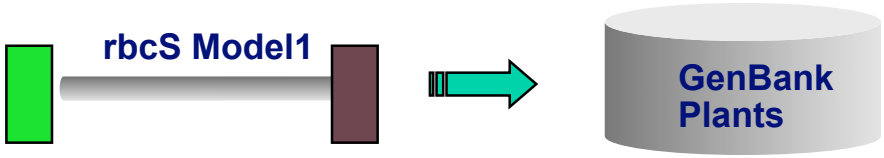
- 9 out of 37 EPD matches were plant promoters:
- 8 rbcS genes (RuBPCss)
- 1 chalcone synthase (also known to be light regulated)

| Sequence | Model Name | Position | Strand | Model Score | Select Match |
|---|---|---|---|---|---|
| GM_RBS4 [EP16010] (1 - 600) [DNA] Gm RuBPCss SRS4; range -499 to 100. | rbcS Model1 | 272 - 309 | (+) | 93.5 % | ☐ |
| PS_RBS3 [EP16012] (1 - 600) [DNA] Ps RuBPCss 3A; range -499 to 100. | rbcS Model1 | 259 - 302 | (+) | 86.7 % | ☐ |
| PS_RBS2_B [EP16013] (1 - 600) [DNA] Ps RuBPCss 3C; range -499 to 100. | rbcS Model1 | 300 - 335 | (+) | 92.9 % | ☐ |
| PH_RBS1 [EP14009] (1 - 600) [DNA] Ph RuBPCss 8/301; range -499 to 100. | rbcS Model1 | 250 - 291 | (+) | 95.5 % | ☐ |
| PH_RBS6 [EP23012] (1 - 600) [DNA] Ph RuBPCss 611; range -499 to 100. | rbcS Model1 | 281 - 323 | (+) | 95.9 % | ☐ |
| NT_RBS [EP16009] (1 - 600) [DNA] Nt RuBPCss Ntss23; range -499 to 100. | rbcS Model1 | 247 - 287 | (+) | 95.4 % | ☐ |
| NT_RBS8 [EP32007] (1 - 600) [DNA] Nt RuBPCss 3-8; range -499 to 100. | rbcS Model1 | 240 - 283 | (+) | 97.5 % | ☐ |
| NP_RBS8 [EP24004] (1 - 600) [DNA] Np RuBPCss 8B; range -499 to 100. | rbcS Model1 | 243 - 286 | (+) | 97.0 % | ☐ |
| AM_CHSY [EP11005] (1 - 600) [DNA] Am chalcone synthase; range -499 to 100. | rbcS Model1 | 341 - 381 | (+) | 94.0 % | ☐ |

Model1 is selective for rbcS and other light regulated genes

# Genomatix Database search using rbcS Model1

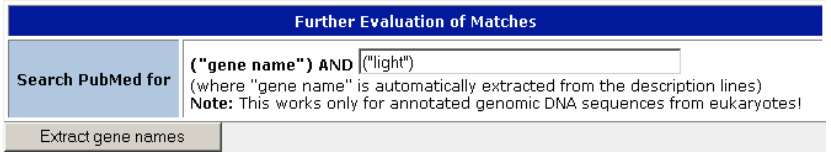
**GEMS Launcher Task: *ModelInspector*: Search for user-defined models, working on Plants (GenBank Release 141, a total of 1.14 Billion bp)**




**A) Result: a total of 9440 matches was found.**
**- A two elements model usually also has spurious matches, outside of the right promoter context.**

**B) => Matches filtered for: promoters => A total of 121 matches was found.**

- 34 matches had a single gene annotation in the description line. For these, a software supported screen for keywords within PubMed was performed:




- 11 were annotated as rbcS
- At least 5 matches are genes also supposed to be light regulated:

phenylalanine ammonia-lyase
chalcone synthase
sucrose-phosphate synthase
nitrate reductase
phosphoglycerate kinase

Model1 is selective for rbcS and other light regulated genes

# Genomatix — Plant module library

**GEMS Launcher Task: *ModelInspector*: Search for promoter modules, working on rbcS_promoters.embl (5 seq.)**

- Library: Module Library Version 3.5, Plant Modules (January 2004)
- ~110 plant matrices
- ~30 plant modules

| Sequence | Model Name | Position | Strand | Model Score | Select Match |
|---|---|---|---|---|---|
| ATRBCSA [X13611] (1 - 701)<br>[DNA] Arabidopsis thaliana ats1A gene for ribulose 1.5-biphoshate carboxylase small subunit (EC 4.1.1.39). | IBOX_BZIP_01 | 413 - 456 | (+) | 99.2 % | ☐ |
| SOYRUBPA [M16889] (1 - 501)<br>[DNA] Soybean ribulose 1,5-bisphosphate carboxylase small subunit (SRS4) gene, complete cds. | IBOX_BZIP_01 | 268 - 309 | (+) | 98.2 % | ☐ |
| TOBRBCS8B [M36685] (1 - 502)<br>[DNA] N.plumbaginifolia ribulose bisphosphate carboxylase (rbcS-8B) gene, complete cds. | IBOX_BZIP_01 | 245 - 287 | (+) | 98.3 % | ☐ |
| NTRUBSS [X02353] (1 - 701)<br>[DNA] Tobacco gene for ribulose 1,5-bisphosphate carboxylase small subunit. | IBOX_BZIP_01 | 385 - 424 | (+) | 98.3 % | ☐ |
| PEARPC [M21356] (1 - 412)<br>[DNA] Pea (P.sativum) small subunit ribulose bisphosphate carboxylase (rbcS-3A) gene, promoter region. | GTBX_GTBX_01 | 260 - 303 | (+) | 84.5 % | ☐ |

| | | |
|---|---|---|
| Origin | Reference | Baum et al., Plant J. 12, 463-469, 1997 (MEDLINE: 9301095); Martinez-Hernández et al., Plant Physiol. 128, 1223-1233, 2002 (MEDLINE: 11950971) |
| | Gene | Tomato ribulose-1,5-bisphosphate carboxylase, small subunit (RBCS2); Tobacco ribulose bisphosphate carboxylase (rbcS-8B). |
| Function | The minimal light-responsive unit of the rbcS promoter consists of an I-box and a G-box element. | |

- Model IBOX_BZIP_01: P$IBOX, P$BZIP, Distance: 19-27 bp
- Model rbcS Model1 : P$IBOX, P$ABRE, Distance: 14-25 bp

Model IBOX_BZIP_01 from Genomatix library is very similar to the *in silico* generated rbcS Model1

Genomatix

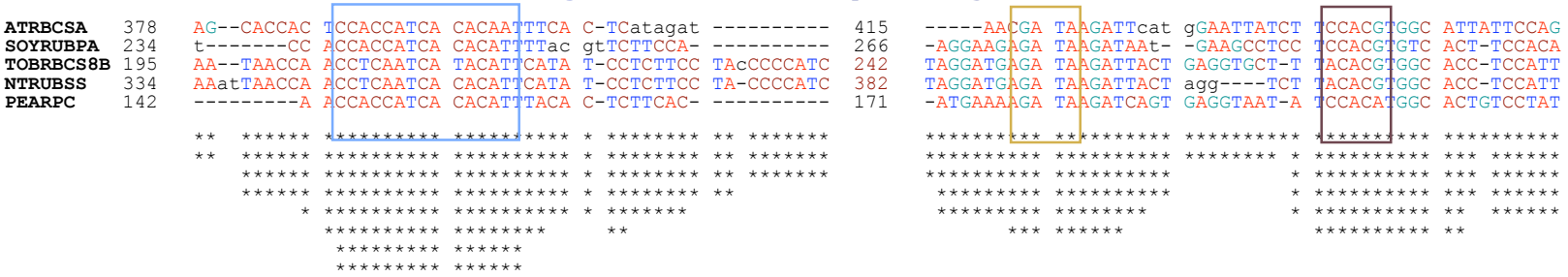
# Task 5

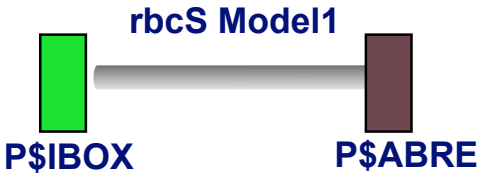
## What you will be doing next

- Analyze rbcS promoters for transcription factor binding sites (TFBS) (***MatInspector*** from ***GEMS*** package)
- Analyze rbcS promoters for TFBS frameworks (***FrameWorker*** from ***GEMS*** package)
- Modify and optimize rbcS promoter models for TFBS frameworks (***FastM*** from ***GEMS*** package)
- Database search with frameworks (***ModelInspector*** from ***GEMS*** package)
- Determine and define new TFBS matrix from multiple alignment (***DiAlign*** and ***MatDefine*** from ***GEMS*** package)
- Experimental design: specifically deleting a TFBS (***SequenceShaper*** from ***GEMS*** package)

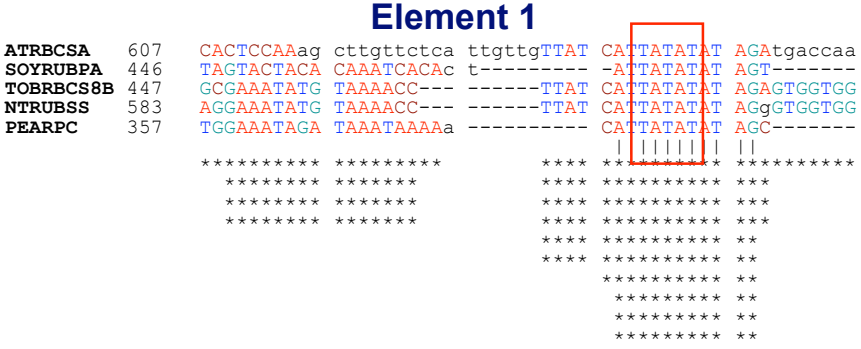
# Sequence alignment

## GEMS Launcher Task: *DiAlign*: Local multiple alignment



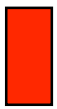




**Element 1**




Four conserved elements:

- Element 1 (unknown)
- IBOX (GATAA), -> light regulation
- ABRE, abscisic acid response. Contains G-Box core (CACGTG), -> light regulation
- TATA-Box



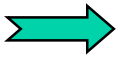
**TATA-Box**

# Definition of weight matrix

## GEMS Launcher Task: *MatDefine*: Definition of weight matrices

- User defined weight matrix for rbcS Element1


Element 1

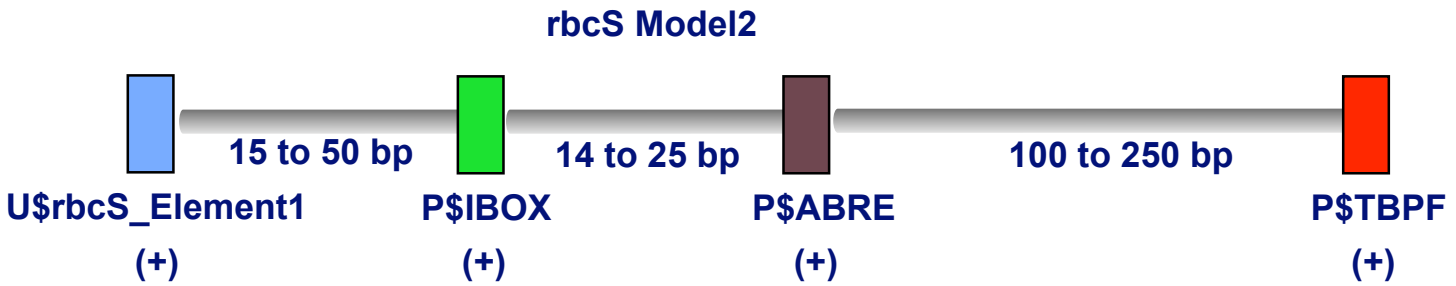



User defined Matrix: U$rbcS_Element1

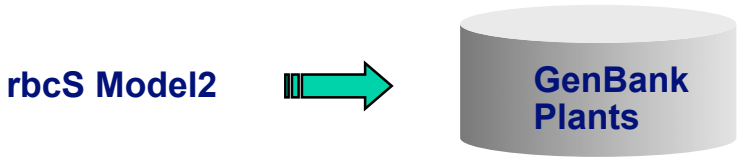
# Enhancement of the rbcS model

**GEMS Launcher Task: Modification of user-defined models**

- Enhance the rbcS model based on matches to the five rbcS promoter training sequences:



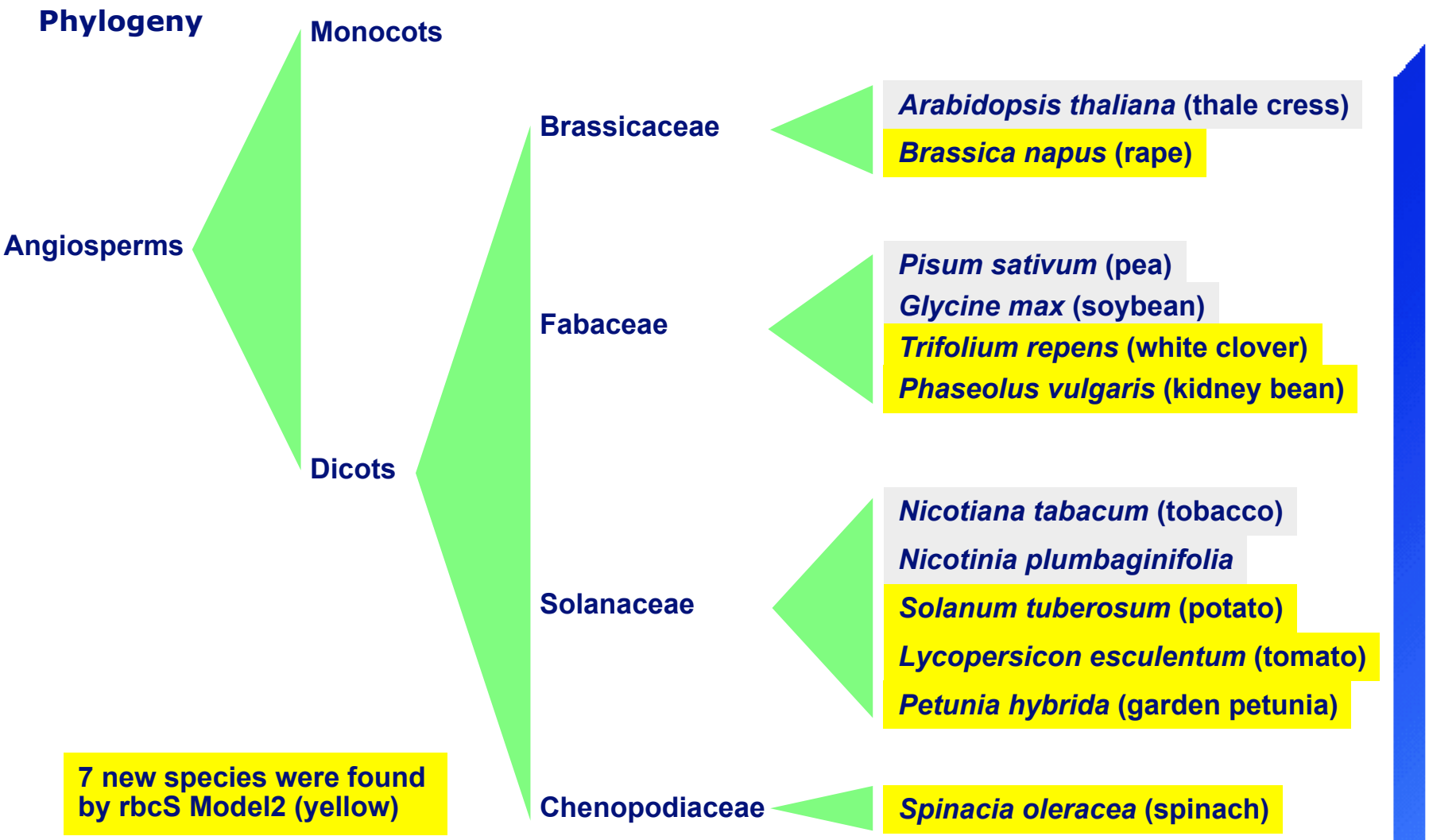

- Database search:




- - 26 matches were found in GenBank plants section (1.14 Billion bp):
    - 22 rbcS genes from different plant species
    - 4 unannotated BAC clones

rbcS Model2 is specific for rbcS gene promoters

# Phylogeny of rbcS Model 2 matches



Genomatix

# Task 6

## What you will be doing next

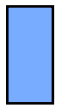
- Analyze rbcS promoters for transcription factor binding sites (TFBS) (***MatInspector*** from ***GEMS*** package)
- Analyze rbcS promoters for TFBS frameworks (***FrameWorker*** from ***GEMS*** package)
- Modify and optimize rbcS promoter models for TFBS frameworks (***FastM*** from ***GEMS*** package)
- Database search with frameworks (***ModelInspector*** from ***GEMS*** package)
- Determine and define new TFBS matrix from multiple alignment (***DiAlign*** and ***MatDefine*** from ***GEMS*** package)
- Experimental design: specifically deleting a TFBS (***SequenceShaper*** from ***GEMS*** package)

# Experimental design

## Pitfalls in mutational analyses

- Is there experimental evidence for the functionality of unknown Element1 ?

```
ATRBCSA     CCACCATCA CACAAT
SOYRUBPA    CCACCATCA CACATT
TOBRBCS8B   CCTCAATCA TACATT
NTRUBSS     CCTCAATCA CACATT
PEARPC      CCACCATCA CACATT
```



**Element 1**

- Reporter assay (wet lab)
- These unintended side effects can happen without experimental design:

***Arabidopsis*:** ATRBCSA AGCACCACTCCACCACCATCACACAATTTCACTCATAGAT

**Deletion:**

AGCACCACTCCACCACCATCACACAATTTCACTCATAGAT

AGCACCACTCCATCACTCATAGAT

**Generates TF binding site:**

| | |
|---|---|
| P$PGCN/GCN4.01 | GCN4, conserved in cereal seed storage protein gene promoters, similar to yeast GCN4 and vertebrate AP-1 |

**Mutation:**

CCACCATCACACAAT

CCACCgaCACACAAT

**Generates TF binding site:**

| | |
|---|---|
| P$CE1F/ABI4.01 | ABA insensitive protein 4 (ABI4) (AtABI4 - Arabidopsis, ZmABI4 - maize, OsABI4 - rice) |

# Experimental design

**GEMS Launcher Task: *SequenceShaper*: Design of regulatory sequences , working on rbcS_At_Element1_context.fasta (40 bp)**

- Experimental design: Specific generation/deletion of TF binding sites without changing the sequence context.

**Matrix libraries: plants.lib, user_defined.lib**

| | Family/matrix | Further Information | Opt. | Position from - to | Str. | Core sim. | Matrix sim. | Sequence |
|---|---|---|---|---|---|---|---|---|
| | U$rbcS_El1/rbcS_Element1 | rbcS Element 1 | 0.71 | 13 - 27 | (+) | 1.000 | 0.967 | ccaccATCAcacaat |

Generate Site | Delete Sites

Max. number of mutations per site: 2

☑ don't delete any other site

☑ don't generate additional site

**ATRBCSA** AGCACCACTCCACCACCATCACACAATTTCACTCATAGAT

| Suggested Mutations | |
|---|---|
| original Sequence | agcaccactcca ccaccatcacacaat ttcactcatagat |
| ☐ (A18C) (T19A) | agcaccactcca ccaccacacacaat ttcactcatagat |
| ☐ (A18C) (T19G) | agcaccactcca ccacccgcacacaat ttcactcatagat |
| ☐ (A18T) (T19G) | agcaccactcca ccacctgcacacaat ttcactcatagat |
| ☐ (A18T) (C20A) | agcaccactcca ccaccttaacacaat ttcactcatagat |

*In silico* sequence design is a prerequisite for wet lab mutational analyses

Genomatix

# Plant references

## Plant references


- Arguello-Astorga GR, Herrera-Estrella LR.
  Ancestral multipartite units in light-responsive plant promoters have structural features correlating with specific phototransduction pathways.
  *Plant Physiol. 1996 Nov; 112(3): 1151-66.*

- Martinez-Hernandez A, Lopez-Ochoa L, Arguello-Astorga G, Herrera-Estrella L.
  Functional properties and regulatory complexity of a minimal RBCS light-responsive unit activated by phytochrome, cryptochrome, and plastid signals.
  *Plant Physiol. 2002 Apr; 128(4): 1223-33.*