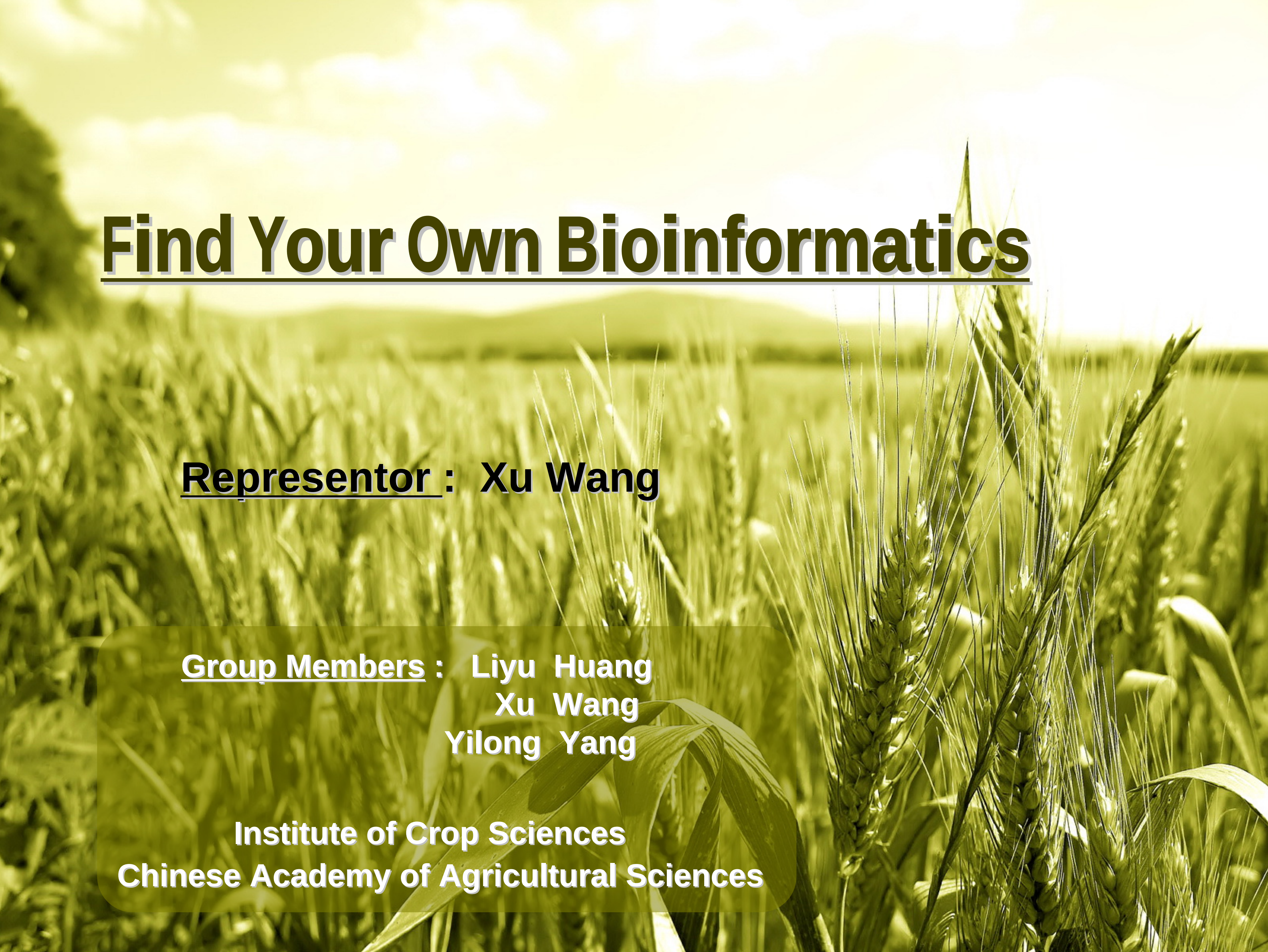


Find Your Own Bioinformatics

Representor : Xu Wang

Group Members : Liyu Huang
Xu Wang
Yilong Yang

Institute of Crop Sciences
Chinese Academy of Agricultural Sciences

What’s for my own ‘Bioinformatics’ ??

- Text mining- finding the nuggets in the literature
  - iHOP
  - GOPubMed

- Before my starting to clone genes …
  - GENEVESTIGATOR
  - Diurnal
  - Codontree

…

- Bioinformatics in Plant Biology

From : Annu. Rev. Plant Biol. 2006. 57:335–60

Text mining- finding the nuggets in the literature

Q : What can text-mining offer us ??

✓ 'The goal of text mining is to allow researchers to identify needed information and shift the burden of searching from researchers to the computer.' *(Rhee, 2006)*

Q : What do I use text-mining tools for ??

✓ searching my interested paper with more ease

✓ extracting the useful data from PubMed

✓ …

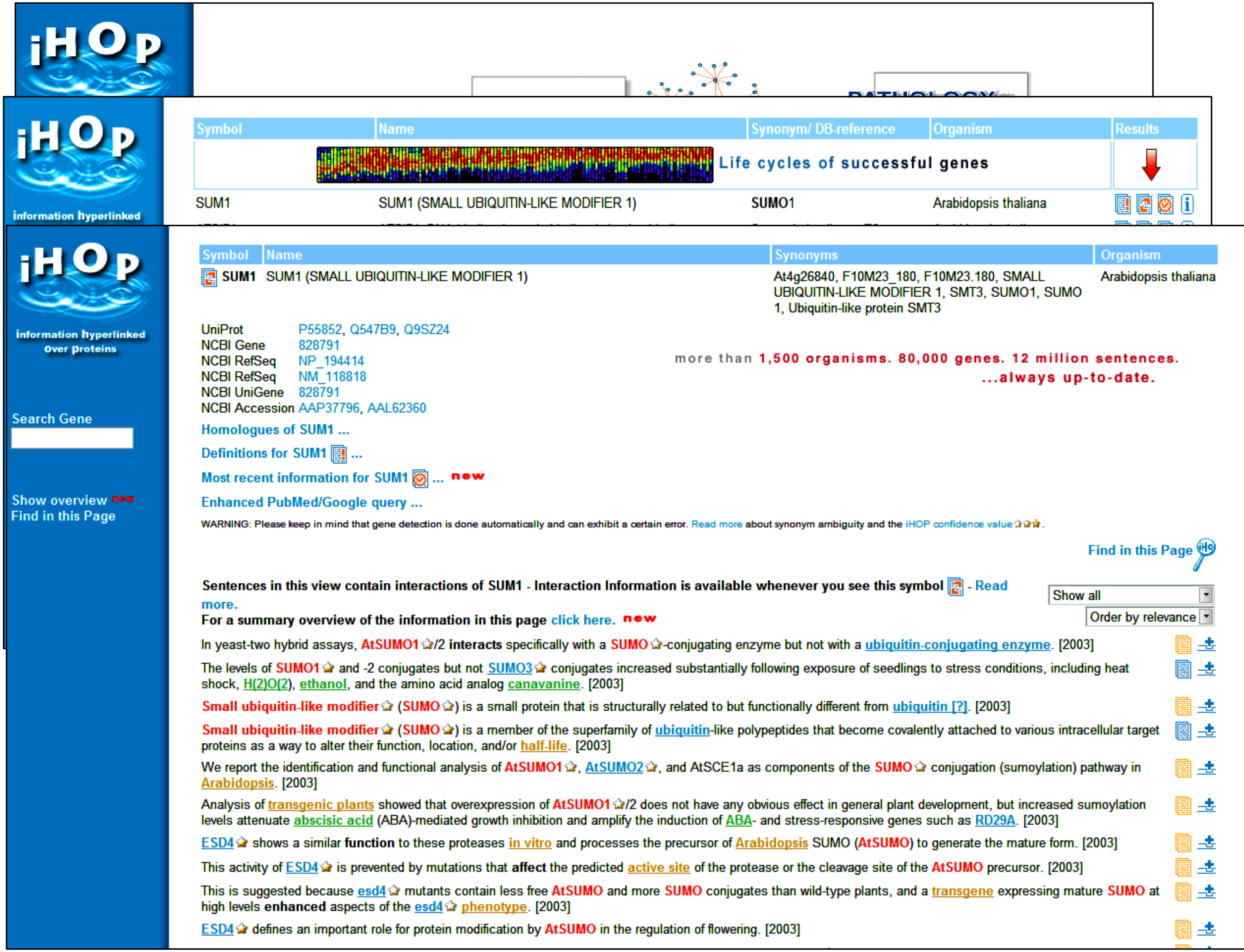
Text mining- finding the nuggets in the literature

Tools I often used：

- iHOP- a gene network for navigating the literature

✓ iHOP provieds the network as a natural way of accessing millions of PubMed abstracts. By using genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed can be converted into one navigable resource, bring all advantages of the Internet to scientiifc literature research.

http://www.ihop-net.org/UniPub/iHOP/



iHOP

iHOP

information hyperlinked

Symbol	Name	Synonym/ DB-reference	Organism	Results
Life cycles of successful genes				
SUM1	SUM1 (SMALL UBIQUITIN-LIKE MODIFIER 1)	SUMO1	Arabidopsis thaliana	

iHOP

information hyperlinked over proteins

Search Gene

Show overview new
Find in this Page

Symbol	Name	Synonyms	Organism
SUM1	SUM1 (SMALL UBIQUITIN-LIKE MODIFIER 1)	At4g26840, F10M23_180, F10M23.180, SMALL UBIQUITIN-LIKE MODIFIER 1, SMT3, SUMO1, SUMO 1, Ubiquitin-like protein SMT3	Arabidopsis thaliana

UniProt	P55852, Q547B9, Q9SZ24
NCBI Gene	828791
NCBI RefSeq	NP_194414
NCBI RefSeq	NM_118818
NCBI UniGene	828791
NCBI Accession	AAP37796, AAL62360

more than 1,500 organisms. 80,000 genes. 12 million sentences.
...always up-to-date.

Homologues of SUM1 ...

Definitions for SUM1 ...

Most recent information for SUM1 ... new

Enhanced PubMed/Google query ...

WARNING: Please keep in mind that gene detection is done automatically and can exhibit a certain error. Read more about synonym ambiguity and the iHOP confidence value.

Find in this Page

Sentences in this view contain interactions of SUM1 - Interaction Information is available whenever you see this symbol - Read more.

For a summary overview of the information in this page click here. new

Show all
Order by relevance

In yeast-two hybrid assays, AtSUMO1/2 **interacts** specifically with a SUMO-conjugating enzyme but not with a ubiquitin-conjugating enzyme. [2003]

The levels of SUMO1 and -2 conjugates but not SUMO3 conjugates increased substantially following exposure of seedlings to stress conditions, including heat shock, H(2)O(2), ethanol, and the amino acid analog canavanine. [2003]

Small ubiquitin-like modifier (SUMO) is a small protein that is structurally related to but functionally different from ubiquitin [?]. [2003]

Small ubiquitin-like modifier (SUMO) is a member of the superfamily of ubiquitin-like polypeptides that become covalently attached to various intracellular target proteins as a way to alter their function, location, and/or half-life. [2003]

We report the identification and functional analysis of AtSUMO1, AtSUMO2, and AtSCE1a as components of the SUMO conjugation (sumoylation) pathway in Arabidopsis. [2003]

Analysis of transgenic plants showed that overexpression of AtSUMO1/2 does not have any obvious effect in general plant development, but increased sumoylation levels attenuate abscisic acid (ABA)-mediated growth inhibition and amplify the induction of ABA- and stress-responsive genes such as RD29A. [2003]

ESD4 shows a similar **function** to these proteases in vitro and processes the precursor of Arabidopsis SUMO (AtSUMO) to generate the mature form. [2003]

This activity of ESD4 is prevented by mutations that **affect** the predicted active site of the protease or the cleavage site of the AtSUMO precursor. [2003]

This is suggested because esd4 mutants contain less free AtSUMO and more SUMO conjugates than wild-type plants, and a transgene expressing mature SUMO at high levels **enhanced** aspects of the esd4 phenotype. [2003]

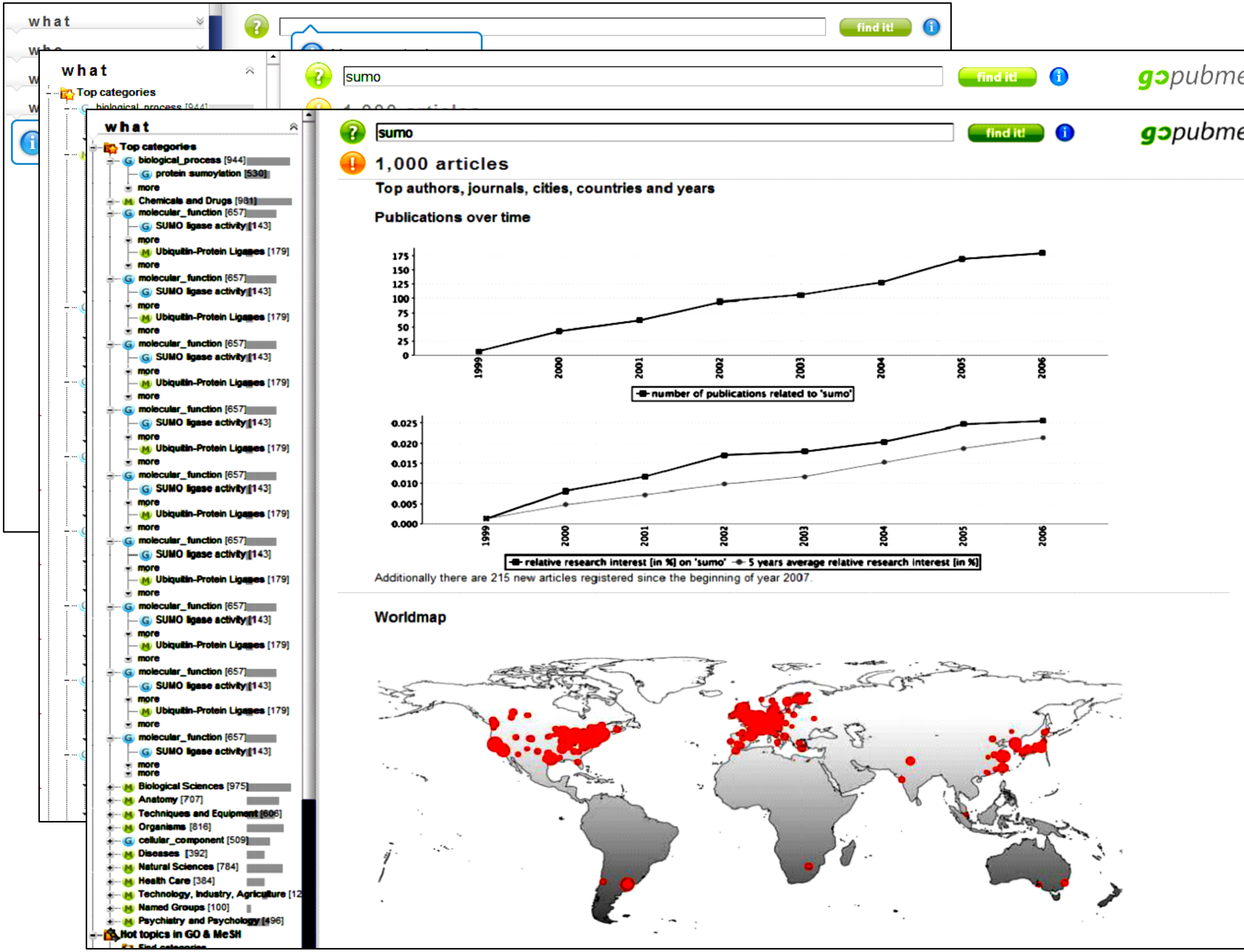
ESD4 defines an important role for protein modification by AtSUMO in the regulation of flowering. [2003]

Text mining- finding the nuggets in the literature

Tools I often used：

- GOPubMed- Gene Ontology and PubMed

http://www.gopubmed.org/



what

- Top categories
  - biological_process [944]
    - protein sumoylation [530]
    - more
  - Chemicals and Drugs [981]
  - molecular_function [657]
    - SUMO ligase activity [143]
    - more
    - Ubiquitin-Protein Ligases [179]
    - more
  - molecular_function [657]
    - SUMO ligase activity [143]
    - more
    - Ubiquitin-Protein Ligases [179]
    - more
  - molecular_function [657]
    - SUMO ligase activity [143]
    - more
    - Ubiquitin-Protein Ligases [179]
    - more
  - molecular_function [657]
    - SUMO ligase activity [143]
    - more
    - Ubiquitin-Protein Ligases [179]
    - more
  - molecular_function [657]
    - SUMO ligase activity [143]
    - more
    - Ubiquitin-Protein Ligases [179]
    - more
  - molecular_function [657]
    - SUMO ligase activity [143]
    - more
    - Ubiquitin-Protein Ligases [179]
    - more
  - molecular_function [657]
    - SUMO ligase activity [143]
    - more
    - Ubiquitin-Protein Ligases [179]
    - more
  - molecular_function [657]
    - SUMO ligase activity [143]
    - more
    - Ubiquitin-Protein Ligases [179]
    - more
  - molecular_function [657]
    - SUMO ligase activity [143]
    - more
    - more
  - Biological Sciences [975]
  - Anatomy [707]
  - Techniques and Equipment [606]
  - Organisms [816]
  - cellular_component [509]
  - Diseases [392]
  - Natural Sciences [784]
  - Health Care [384]
  - Technology, Industry, Agriculture [12
  - Named Groups [100]
  - Psychiatry and Psychology [496]
- Hot topics in GO & MeSH
  - Find categories

sumo

1,000 articles

Top authors, journals, cities, countries and years

Publications over time

175
150
125
100
75
50
25
0
1999
2000
2001
2002
2003
2004
2005
2006
number of publications related to 'sumo'
0.025
0.020
0.015
0.010
0.005
0.000
1999
2000
2001
2002
2003
2004
2005
2006
relative research interest [in %] on 'sumo'
5 years average relative research interest [in %]

Additionally there are 215 new articles registered since the beginning of year 2007.

Worldmap

Before my starting to clone genes …

Some helpful considerations :

- ✓ Is your gene expressed with tissue-specificity ??
- ✓ Is your gene expressed throughout the life cycle ??
- ✓ How to determine the sampling time within a day ??
- ✓ How to refine the degenerate primers in homological cloning??
- ✓ …

tissue-specific

developmental stage-specific

photoperiod

circadian clock

Before my starting to clone genes …

Useful tools based on microarrays data

- GENEVESTIGATOR

  Estimate the tissue and developmental stage specificity

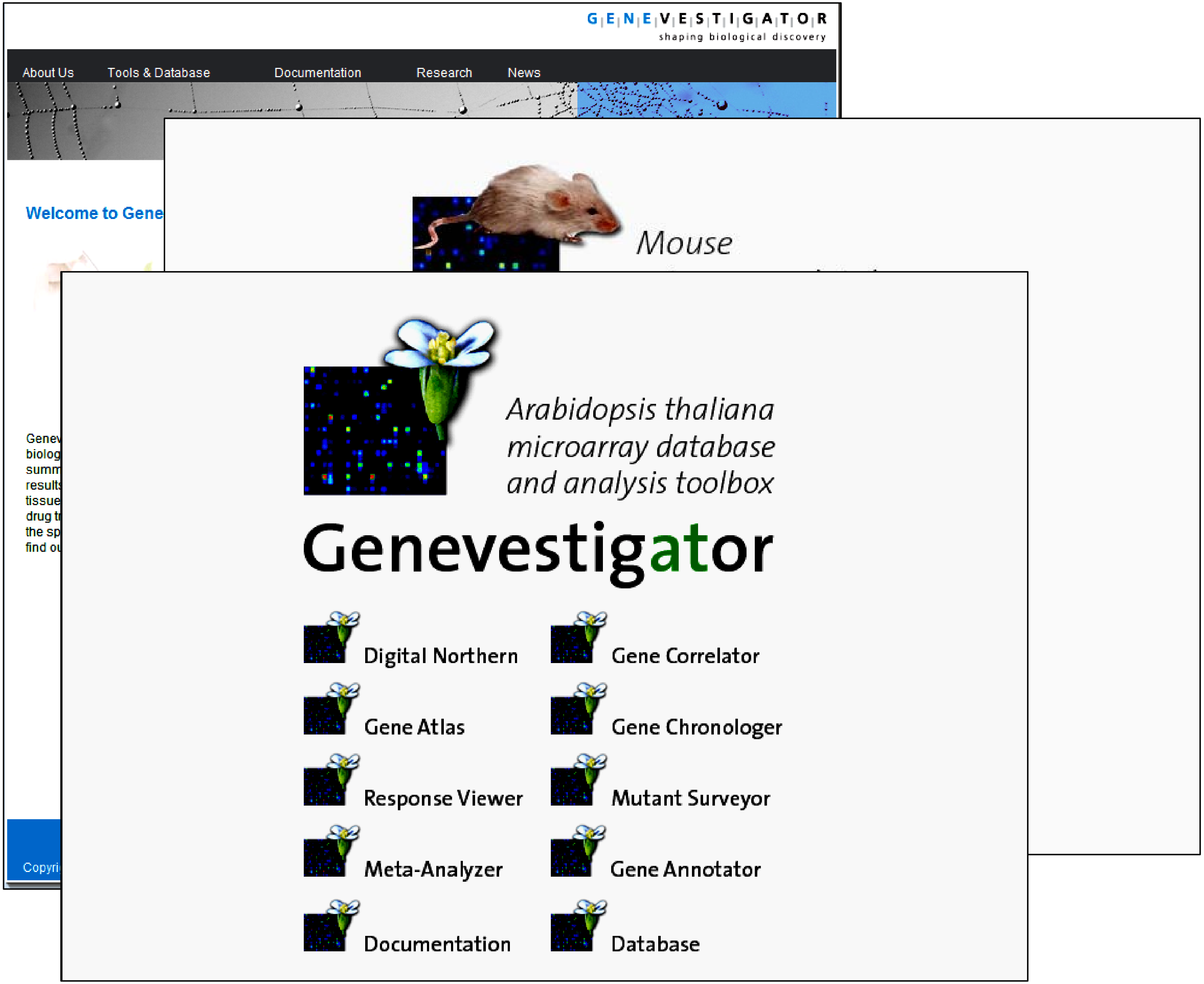
  https://www.genevestigator.ethz.ch/

- Diurnal : only for Arabidopsis genes

  Estimate diurnal and circadian gene expression profile

  http://diurnal.cgrb.oregonstate.edu/

Example: clone the *4CL1* gene in Arabidopsis



Welcome to Gene

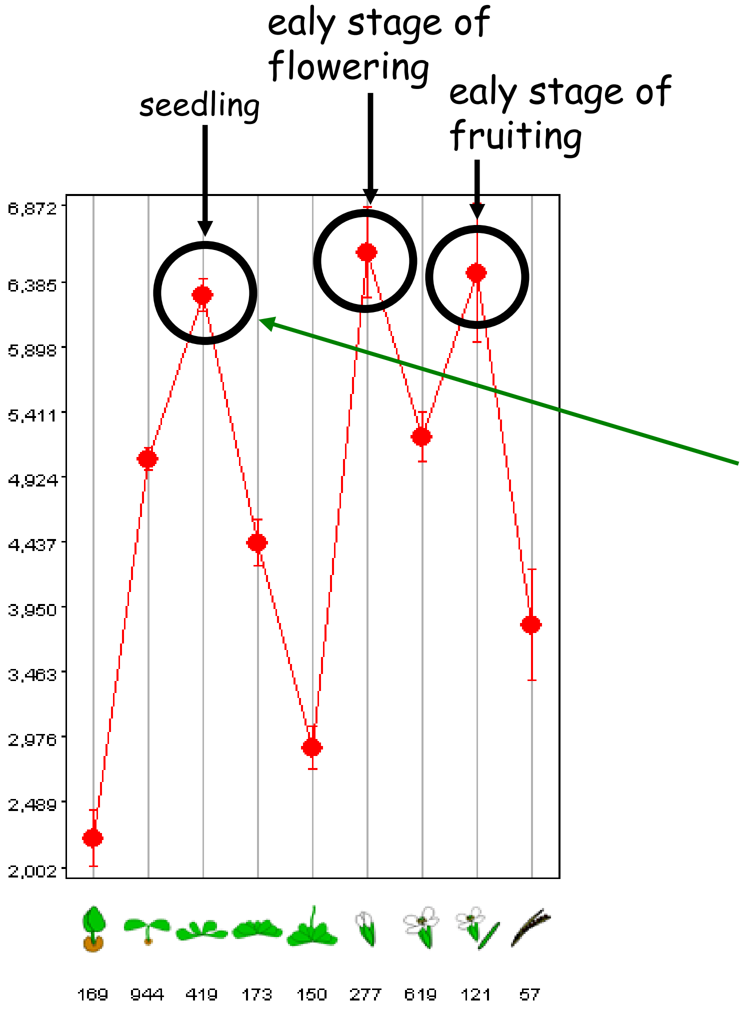
Genev
biolog
summ
results
tissue
drug tr
the sp
find ou

Copyri

Arabidopsis thaliana microarray database and analysis toolbox

Genevestigator

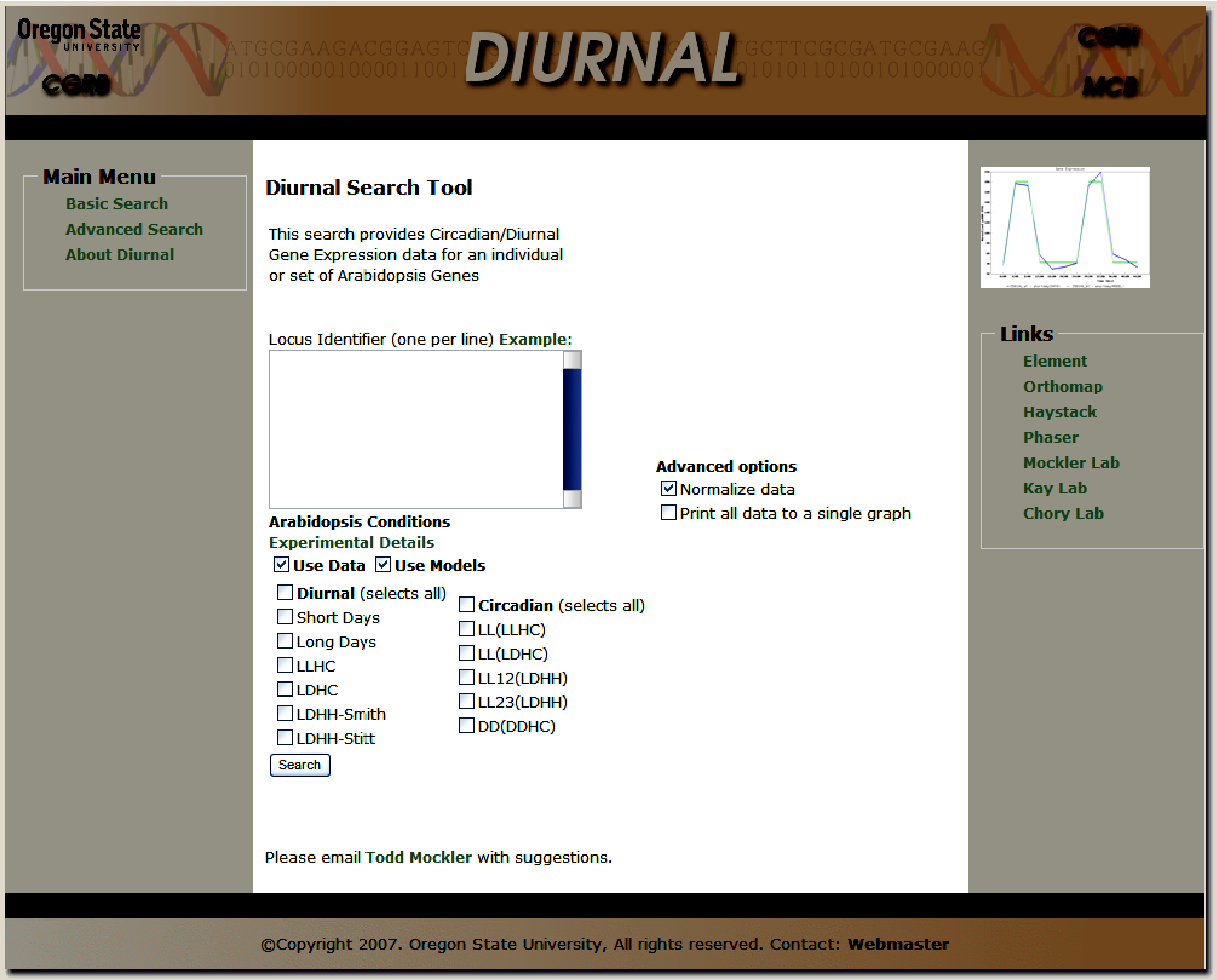
Before my starting to clone genes …



Example: clone the *4CL1* gene in Arabidopsis

What I used:

the whole plant of seedlings without tissue-bias



Main Menu

Basic Search
Advanced Search
About Diurnal

Diurnal Search Tool

This search provides Circadian/Diurnal Gene Expression data for an individual or set of Arabidopsis Genes

Locus Identifier (one per line) **Example:**

Arabidopsis Conditions
Experimental Details

☑ **Use Data** ☑ **Use Models**

□ **Diurnal** (selects all)
□ Short Days
□ Long Days
□ LLHC
□ LDHC
□ LDHH-Smith
□ LDHH-Stitt

□ **Circadian** (selects all)
□ LL(LLHC)
□ LL(LDHC)
□ LL12(LDHH)
□ LL23(LDHH)
□ DD(DDHC)

Search

Advanced options
☑ Normalize data
□ Print all data to a single graph

Links

Element
Orthomap
Haystack
Phaser
Mockler Lab
Kay Lab
Chory Lab

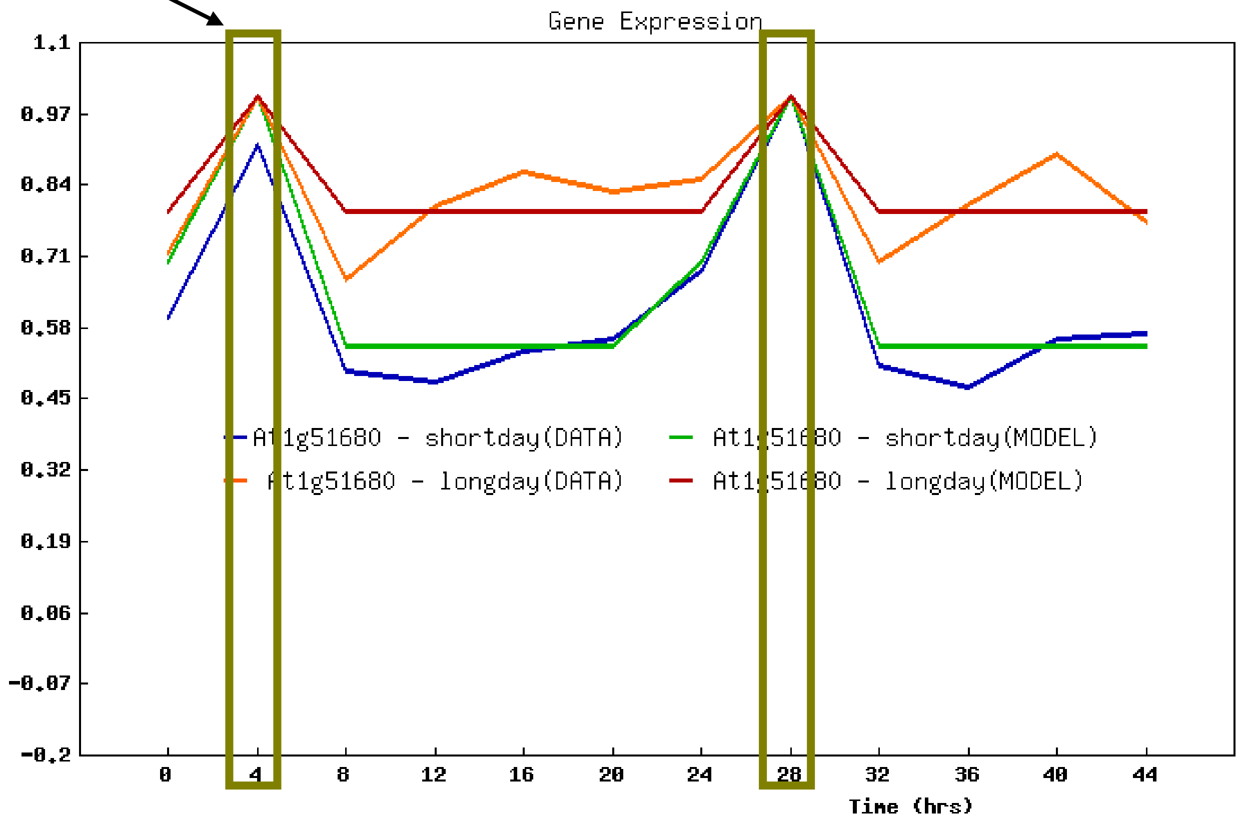
Please email **Todd Mockler** with suggestions.

©Copyright 2007. Oregon State University, All rights reserved. Contact: **Webmaster**

☙ Before my starting to clone genes …

' the expression peak accurs at 4 hours after light '

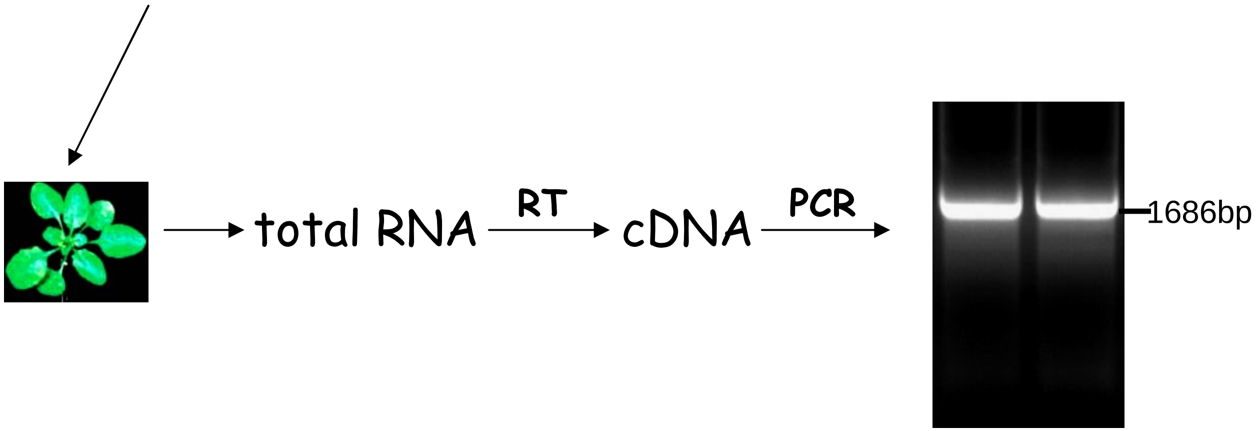
Example: clone the *4CL1* gene in Arabidopsis



Before my starting to clone genes …

Example: clone the *4CL1* gene in Arabidopsis

All those information considered, the sample for mRNA isolation might be the young seedlings collected at 4 hours after the light turned on.



→ total RNA $\xrightarrow{\text{RT}}$ cDNA $\xrightarrow{\text{PCR}}$

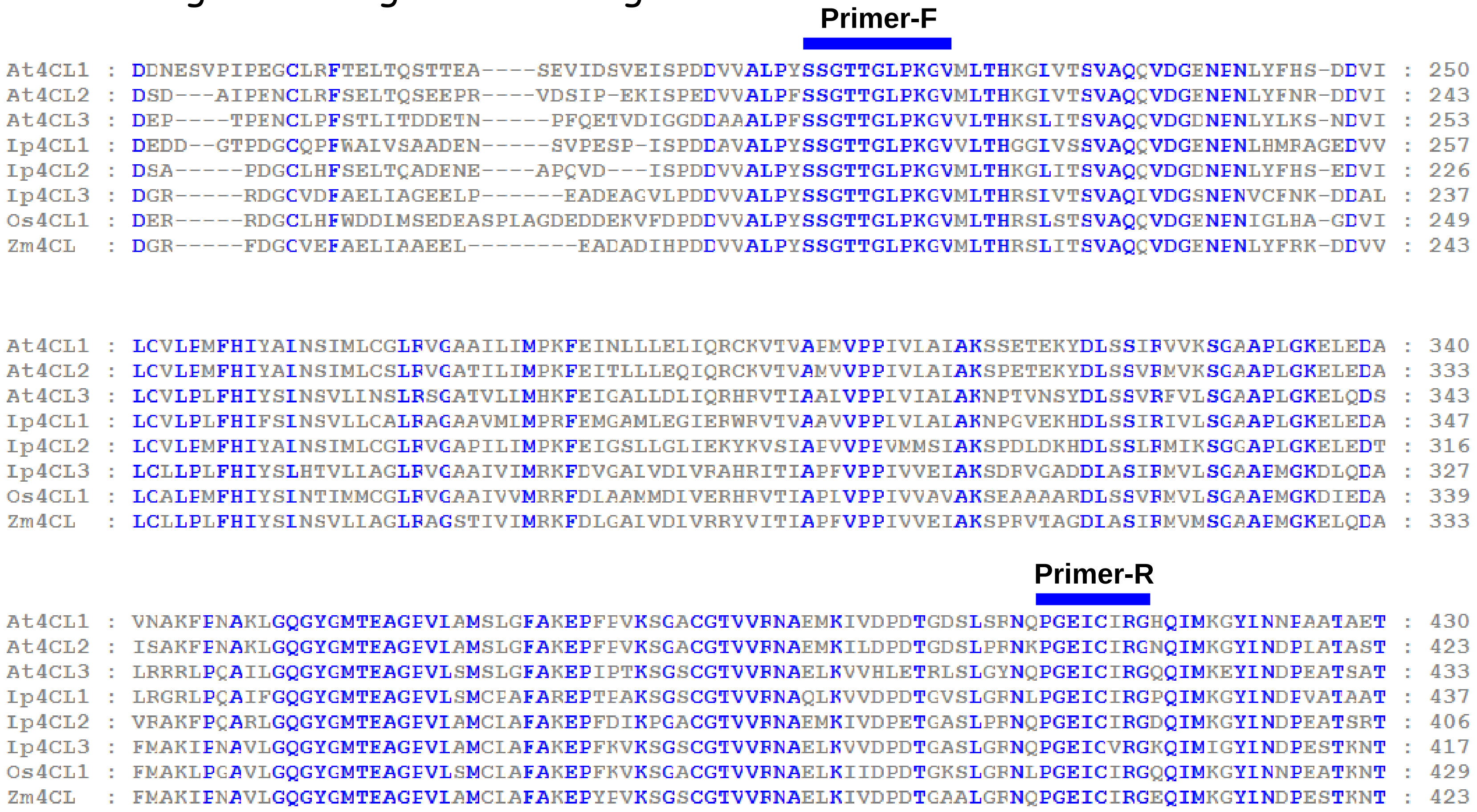
“ Half day on the Web,

saves you half month in the lab! ”

Before my starting to clone genes …

How to refine the degenerate primers for homological cloning ??

- ✓ *4CL* gene cloning from swtichgrass



Primer-F
Primer-R

```
At4CL1 : DDNESVPIPEGCLRFTELTQSTTEA----SEVIDSVEISPDDVVALPYSSGTTGLPKGVMLTHKGIVTSVAQQVDGENPNLYFHS-DDVI : 250
At4CL2 : DSD---AIPENCLRFSELTQSEEPR----VDSIP-EKISPEDVVALPFSSGTTGLPKGVMLTHKGIVTSVAQQVDGENPNLYFNR-DDVI : 243
At4CL3 : DEP----TPENCLPFSTLITDDETN-----PFQETVDIGGDDAAALPFSSGTTGLPKGVVLTHKSLITSVAQQVDGDNPNLYLKS-NDVI : 253
Ip4CL1 : DEDD--GTPDGCQPFWAIVSAADEN-----SVPESP-ISPDDAVALPYSSGTTGLPKGVVLTHGGIVSSVAQQVDGENPNLHMRAGEDVV : 257
Ip4CL2 : DSA-----PDGCLHFSELTQADENE----APQVD---ISPDDVVALPYSSGTTGLPKGVMLTHKGLITSVAQQVDGDNPNLYFHS-EDVI : 226
Ip4CL3 : DGR-----RDGCVDFAELIAGEELP------EADEAGVLPDDVVALPYSSGTTGLPKGVMLTHRSIVTSVAQIVDGSNPNVCFNK-DDAL : 237
Os4CL1 : DER-----RDGCLHFWDDIMSEDEASPLAGDEDDEKVFDPDDVVALPYSSGTTGLPKGVMLTHRSLSTSVAQQVDGENPNIGLHA-GDVI : 249
Zm4CL  : DGR-----FDGCVEFAELIAAEEL--------EADADIHPDDVVALPYSSGTTGLPKGVMLTHRSLITSVAQQVDGENPNLYFRK-DDVV : 243


At4CL1 : LCVLPMFHIYAINSIMLCGLRVGAAILIMPKFEINLLLELIQRCKVTVAPMVPPIVIAIAKSSETEKYDLSSIRVVKSGAAPLGKELEDA : 340
At4CL2 : LCVLPMFHIYAINSIMLCSLRVGATILIMPKFEITLLLEQIQRCKVTVAMVVPPIVIAIAKSPETEKYDLSSVRMVKSGAAPLGKELEDA : 333
At4CL3 : LCVLPLFHIYSINSVLINSLRSGATVLIMHKFEIGALLDLIQRHRVTIAAIVPPIVIALAKNPTVNSYDLSSVRFVLSGAAPLGKELQDS : 343
Ip4CL1 : LCVLPLFHIFSINSVLLCALRAGAAVMIMPRFEMGAMLEGIERWRVTVAAVVPPIVIALAKNPGVEKHDLSSIRIVLSGAAPLGKELEDA : 347
Ip4CL2 : LCVLPMFHIYAINSIMLCGLRVGAPILIMPKFEIGSLLGLIEKYKVSIAPVVPPVMMSIAKSPDLDKHDLSSLRMIKSGGAPLGKELEDT : 316
Ip4CL3 : LCLLPLFHIYSLHTVLIAGLRVGAAIVIMRKFDVGAIVDIVRAHRITIAPFVPPIVVEIAKSDRVGADDIASIRMVLSGAAPMGKDLQDA : 327
Os4CL1 : LCALPMFHIYSINTIMMCGLRVGAAIVVMRRFDIAAMMDIVERHRVTIAPIVPPIVVAVAKSEAAAARDLSSVRMVLSGAAPMGKDIEDA : 339
Zm4CL  : LCLLPLFHIYSINSVLIAGLRAGSTIVIMRKFDLGAIVDIVRRYVITIAPFVPPIVVEIAKSPRVTAGDIASIRMVMSGAAPMGKELQDA : 333


At4CL1 : VNAKFPNAKLGQGYGMTEAGPVIAMSLGFAKEPFPVKSGACGTVVRNAEMKIVDPDTGDSLSRNQPGEICIRGHQIMKGYINNPAATAET : 430
At4CL2 : ISAKFPNAKLGQGYGMTEAGPVIAMSLGFAKEPFPVKSGACGTVVRNAEMKILDPDTGDSLPRNKPGEICIRGNQIMKGYINDPIATAST : 423
At4CL3 : LRRRLPQAILGQGYGMTEAGPVLSMSLGFAKEPIPTKSGSCGTVVRNAELKVVHLETRLSLGYNQPGEICIRGQQIMKEYINDPEATSAT : 433
Ip4CL1 : LRGRLPQAIFGQGYGMTEAGPVLSMCPAFAREPTPAKSGSCGTVVRNAQLKVVDPDTGVSLGRNLPGEICIRGPQIMKGYINDPVATAAT : 437
Ip4CL2 : VRAKFPQARLGQGYGMTEAGPVIAMCIAFAKEPFDIKPGACGTVVRNAEMKIVDPETGASLPRNQPGEICIRGDQIMKGYINDPEATSRT : 406
Ip4CL3 : FMAKIPNAVLGQGYGMTEAGPVIAMCIAFAKEPFKVKSGSCGTVVRNAELKVVDPDTGASLGRNQPGEICVRGKQIMIGYINDPESTKNT : 417
Os4CL1 : FMAKLPGAVLGQGYGMTEAGPVLSMCIAFAKEPFKVKSGACGTVVRNAELKIIDPDTGKSLGRNLPGEICIRGQQIMKGYINNPEATKNT : 429
Zm4CL  : FMAKIPNAVLGQGYGMTEAGPVIAMCIAFAKEPYPVKSGSCGTVVRNAELKIVDPDTGAALGRNQPGEICIRGEQIMKGYINDPESTKNT : 423
```

Before my starting to clone genes …

How to refine the degenerate primers for homological cloning ??

✓ *4CL* gene cloning from swtichgrass

Degenerate primers:

SSGTTGLPKGV

WSNWSNGGNACNACNGGNYTNCCNAARGGNGTN

PGEICIRG

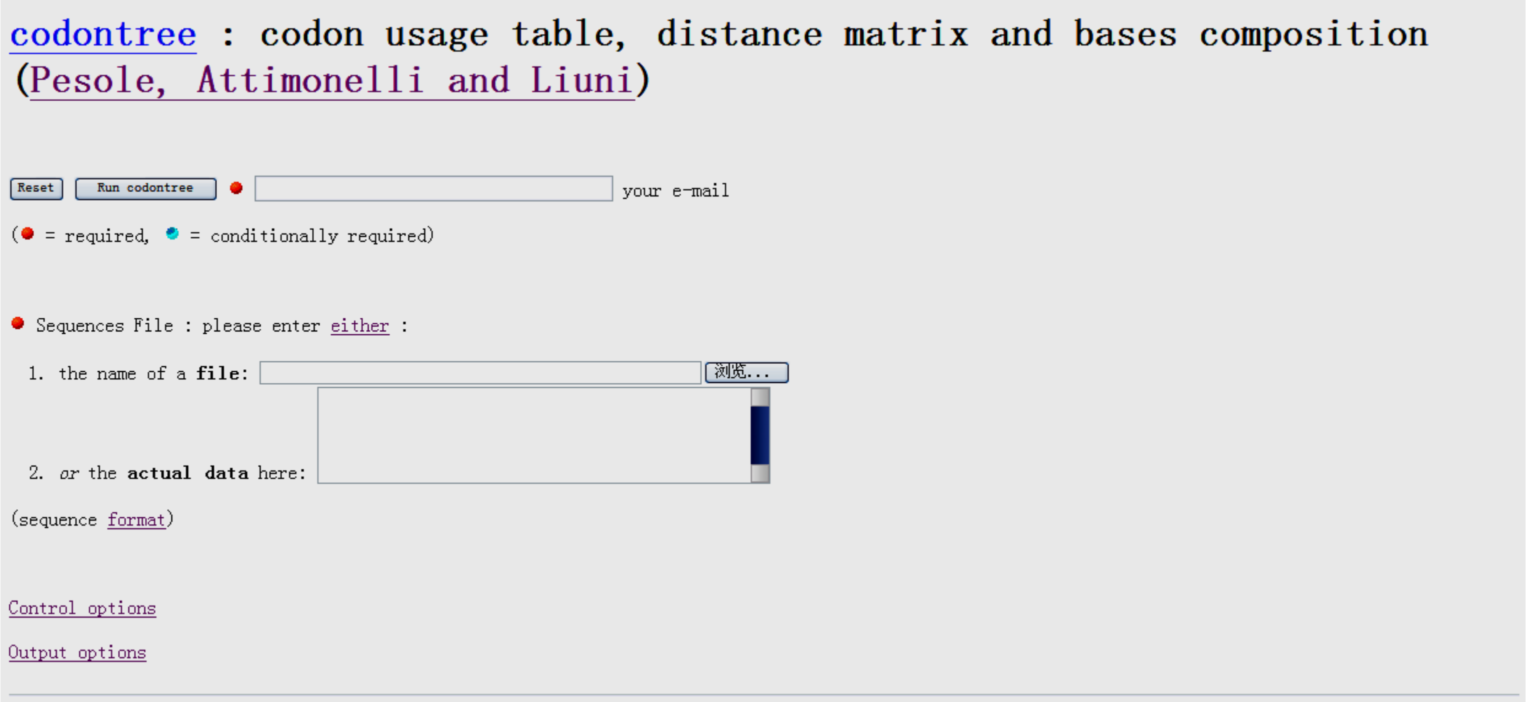
CCNGGNGARATHTGYATHMGNGGN

So many uncertain sites, how to deal with ??

Before my starting to clone genes …

✓ *4CL* gene cloning from swtichgrass

What I did … **Condontree** for looking into condon bias



codontree : codon usage table, distance matrix and bases composition (Pesole, Attimonelli and Liuni)

● your e-mail

(● = required, ● = conditionally required)

● Sequences File : please enter either :

1. the name of a **file**:

2. *or* the **actual data** here:

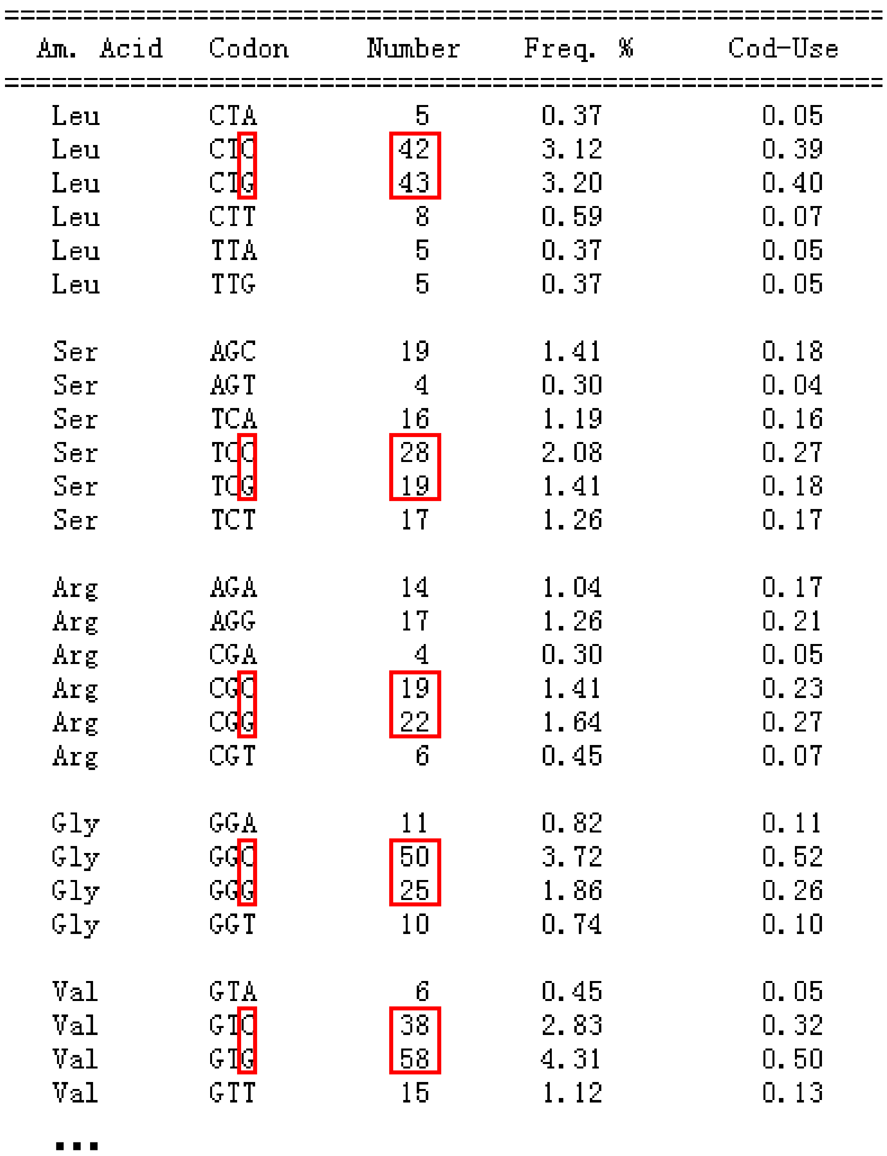
(sequence format)

Control options

Output options

http://bioweb.pasteur.fr/seqanal/interfaces/codontree.html

Before my starting to clone genes …

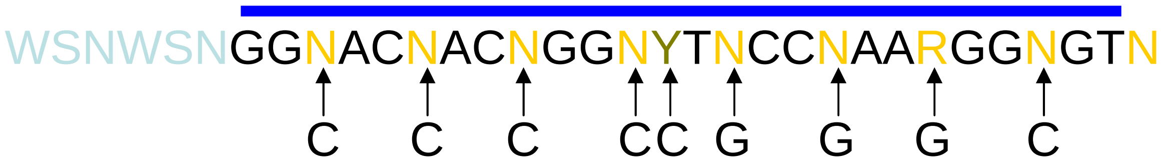


```
=============================================================
  Am. Acid     Codon      Number     Freq. %        Cod-Use
=============================================================
   Leu          CTA           5       0.37            0.05
   Leu          CTC          42       3.12            0.39
   Leu          CTG          43       3.20            0.40
   Leu          CTT           8       0.59            0.07
   Leu          TTA           5       0.37            0.05
   Leu          TTG           5       0.37            0.05

   Ser          AGC          19       1.41            0.18
   Ser          AGT           4       0.30            0.04
   Ser          TCA          16       1.19            0.16
   Ser          TCC          28       2.08            0.27
   Ser          TCG          19       1.41            0.18
   Ser          TCT          17       1.26            0.17

   Arg          AGA          14       1.04            0.17
   Arg          AGG          17       1.26            0.21
   Arg          CGA           4       0.30            0.05
   Arg          CGC          19       1.41            0.23
   Arg          CGG          22       1.64            0.27
   Arg          CGT           6       0.45            0.07

   Gly          GGA          11       0.82            0.11
   Gly          GGC          50       3.72            0.52
   Gly          GGG          25       1.86            0.26
   Gly          GGT          10       0.74            0.10

   Val          GTA           6       0.45            0.05
   Val          GTC          38       2.83            0.32
   Val          GTG          58       4.31            0.50
   Val          GTT          15       1.12            0.13
   ...
```

Before my starting to clone genes …

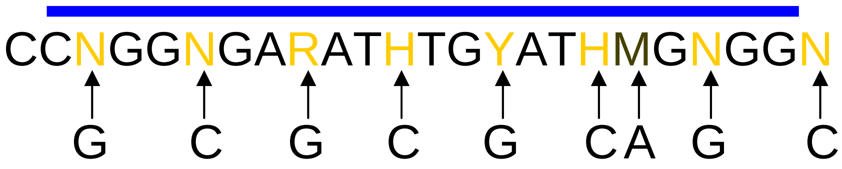
- ✓ *4CL* gene cloning from swtichgrass

primer-F :   5’-CACCACCGGCCTGCCGAAGGGCGT-3’


WSNWSNGGNACNACNGGNYTNCCNAARGGNGTN

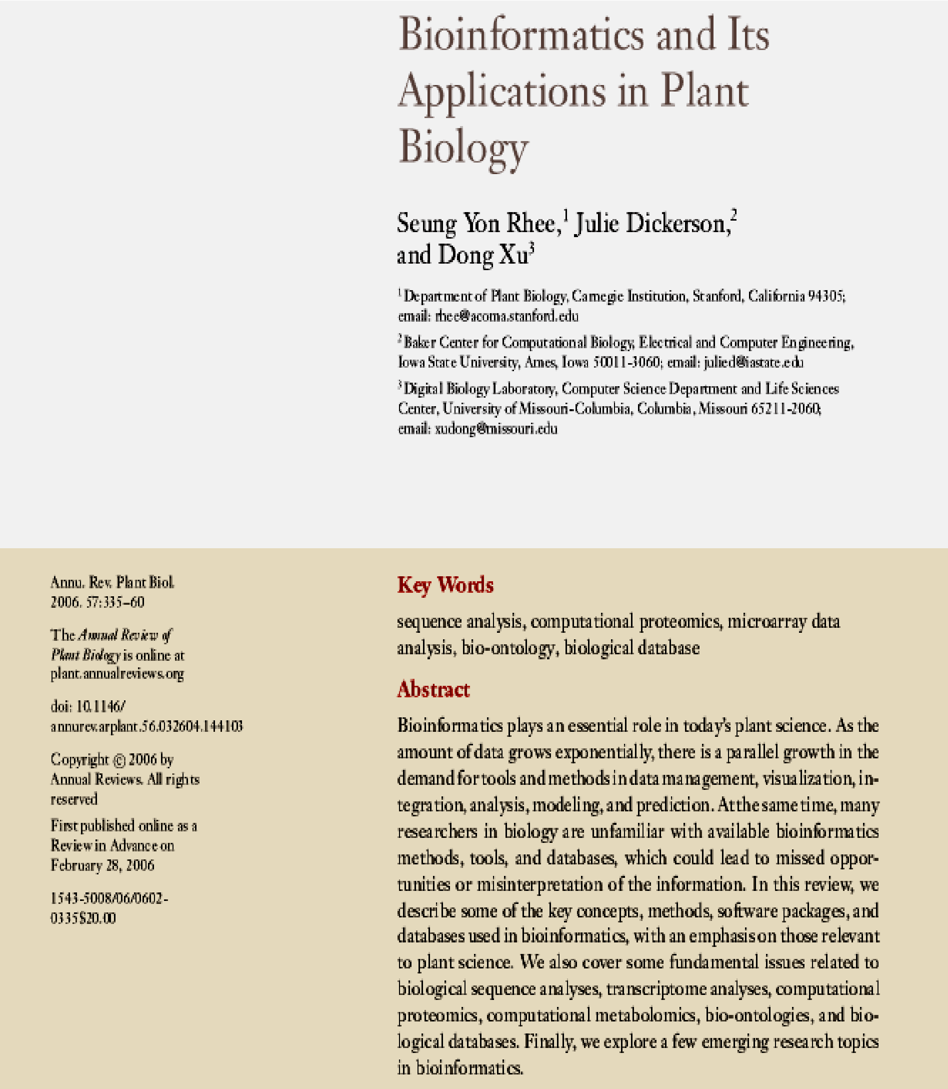
C C C CC G G G C

primer-R :   5’-CGGGCGAGATCTGGATCAGGGG-3’


CCNGGNGARATHTGYATHMGNGGN

G C G C G CA G C

Bioinformatics in Plant Biology



Bioinformatics and Its Applications in Plant Biology

Seung Yon Rhee,[1] Julie Dickerson,[2] and Dong Xu[3]

[1] Department of Plant Biology, Carnegie Institution, Stanford, California 94305; email: rhee@acoma.stanford.edu

[2] Baker Center for Computational Biology, Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011-3060; email: julied@iastate.edu

[3] Digital Biology Laboratory, Computer Science Department and Life Sciences Center, University of Missouri-Columbia, Columbia, Missouri 65211-2060; email: xudong@missouri.edu

Annu. Rev. Plant Biol.
2006. 57:335–60

The *Annual Review of Plant Biology* is online at plant.annualreviews.org

doi: 10.1146/annurev.arplant.56.032604.144103

First published online as a Review in Advance on February 28, 2006

1543-5008/06/0602-0335$20.00

Key Words

sequence analysis, computational proteomics, microarray data analysis, bio-ontology, biological database

Abstract

Bioinformatics plays an essential role in today's plant science. As the amount of data grows exponentially, there is a parallel growth in the demand for tools and methods in data management, visualization, integration, analysis, modeling, and prediction. At the same time, many researchers in biology are unfamiliar with available bioinformatics methods, tools, and databases, which could lead to missed opportunities or misinterpretation of the information. In this review, we describe some of the key concepts, methods, software packages, and databases used in bioinformatics, with an emphasis on those relevant to plant science. We also cover some fundamental issues related to biological sequence analyses, transcriptome analyses, computational proteomics, computational metabolomics, bio-ontologies, and biological databases. Finally, we explore a few emerging research topics in bioinformatics.

Annu. Rev. Plant Biol. 2006. 57:335–60

Acknowledgments

Deeply appreciation to Prof. JC Luo for his precious edifications and instructions !

Cordial thanks to all my classmates who made the whole course study agreeable !

Thanks for
your attentions !