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
  
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 provides 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, bringing all advantages of the Internet to scientific literature research.

http://www.ihop-net.org/UniPub/iHOP/
## Life cycles of successful genes

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
<th>Synonym/DB-reference</th>
<th>Organism</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUM1</td>
<td>SUM1 (SMALL UBIQUITIN-LIKE MODIFIER 1)</td>
<td>SUMO1</td>
<td>Arabidopsis thaliana</td>
<td>![image]</td>
</tr>
</tbody>
</table>

**UniProt** P55852, Q54789, Q93224
**NCBI Gene** 582791
**NCBI RefSeq** NP_194414
**NCBI RefSeq** NM_115616
**NCBI UniGene** 582791
**NCBI Accession** AAP37796, AAL89269

**Homologues of SUM1** ...
**Definitions for SUM1** ...
**Most recent information for SUM1** ...

**Enhanced PubMed/Google query** ...

---

**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

In yeast two-hybrid assays, AtSUMO1Δ2A interacts specifically with a SUMO Δ-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, H2O2, etanol, and the amino acid analog canavanine. [2003]

Small ubiquitin-like modifier 2 (SUMO-2) is a small protein that is structurally related but functionally different from ubiquitin. [2003]

Small ubiquitin-like modifier 2 (SUMO-2) 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 AtSUMO1Δ2A, AtSUMO2Δ2A, and ASC1Δ1 as components of the SUMO Δ-conjugation (sumoylation) pathway in Arabidopsis. [2003]

Analysis of transgenic plants showed that overexpression of AtSUMO1Δ2A 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]

ESD1A 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 ESD1A is prevented by mutations that affect the predicted active site of the proteases or the cleavage sites of the AtSUMO precursors. [2003]

This is suggested because edd1 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 edd1 phenotype. [2003]

ESD1A defines a major 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/
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??
- …
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
Genevestigator

- Digital Northern
- Gene Atlas
- Response Viewer
- Meta-Analyzer
- Documentation

- Gene Correlator
- Gene Chronologer
- Mutant Surveyor
- Gene Annotator
- Database
Before my starting to clone genes ...

Example: clone the 4CL1 gene in Arabidopsis

What I used:
the whole plant of seedlings without tissue-bias
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:

Advanced options
- Normalize data
- Print all data to a single graph

Arabidopsis Conditions

Experimental Details
- Use Data
- Use Models

- Diurnal (selects all)
  - Short Days
  - Long Days
  - LLHC
  - LDHC
  - LDHC-Smith
  - LDHC-Sitt

Search

Please email Todd Mockler with suggestions.
Before my starting to clone genes ...

' the expression peak occurs 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.

“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 switchgrass

Primer-F

Primer-R

At4CL1 : DDNESVP1EPFEGCLRFTELQGSTTEA----SEVIDSVEISPDEVVALFYSSTDGLPKGVMLTHKVFTSVAQCDGENENLYFHS-SDVI : 250
At4CL2 : DSD-----AIPEONCLRFSELTQSEEFR----VDSIP-EKISEPELVVALFYSSTDGLPKGVMLTHKFTSVAQCDGENENLYFNR-SDVI : 243
At4CL3 : DET-----TEENCLRFSTILTDEVEIN-------FFCEETVDIGGDEAAALFYSSTDGLPKGVMLTHKFTSVAQCDGENENLYLKS-NDVI : 253
Lp4CL1 : DEDD----CTEDGCPFWAIWTAADEN-------SVFESP-ISFDPDVALFYSSTDGLPKGVMLTHKFTSVAQCDGENENLMPAGEDDV : 257
Lp4CL2 : DSA----FDGDLRFSELTQGADENE-------AFQVD-----ISDEDFVALFYSSTDGLPKGVMLTHKFTSVAQCDGENENLMPAGEDDV : 226
Lp4CL3 : DGR-----RDGCVDFAIWIAKILEF--------EADEACVDFALFYSSTDGLPKGVMLTHKFTSVAQCDGENENLMPAGEDDV : 237
Os4CL1 : DER-----NDGCLFFHDDIMSEDEASAPAGDEDDENFVDIFIEDVALFYSSTDGLPKGVMLTHKFTSVAQCDGENENLMPAGEDDV : 249
Zm4CL1 : DGR-----FDGCVFVAELIAAEEL--------EADEAHHFALFYSSTDGLPKGVMLTHKFTSVAQCDGENENLMPAGEDDV : 243

At4CL1 : LCVLFMFIYAINILIMCGLFVGCAAILMFKFELNILEL1LQQRCKVTVAFNVVPFVIYAIAKSSEBETYKVLSRITRVKCSAAPMLGKELEDA : 340
At4CL2 : LCVLFMFIYAINILIMCGLFVGCAAILMFKFELNILEL1LQQRCKVTVAFNVVPFVIYAIAKSSEBETYKVLSRITRVKCSAAPMLGKELEDA : 333
At4CL3 : LCVLFMFIYAINILIMCGLFVGCAAILMFKFELNILEL1LQQRCKVTVAFNVVPFVIYAIAKSSEBETYKVLSRITRVKCSAAPMLGKELEDA : 333
Lp4CL1 : LCVLFMFIYAINILIMCGLFVGCAAILMFKFELNILEL1LQQRCKVTVAFNVVPFVIYAIAKSSEBETYKVLSRITRVKCSAAPMLGKELEDA : 347
Lp4CL2 : LCVLFMFIYAINILIMCGLFVGCAAILMFKFELNILEL1LQQRCKVTVAFNVVPFVIYAIAKSSEBETYKVLSRITRVKCSAAPMLGKELEDA : 347
Lp4CL3 : LCVLFMFIYAINILIMCGLFVGCAAILMFKFELNILEL1LQQRCKVTVAFNVVPFVIYAIAKSSEBETYKVLSRITRVKCSAAPMLGKELEDA : 347
Os4CL1 : LCVLFMFIYAINILIMCGLFVGCAAILMFKFELNILEL1LQQRCKVTVAFNVVPFVIYAIAKSSEBETYKVLSRITRVKCSAAPMLGKELEDA : 339
Zm4CL1 : LCVLFMFIYAINILIMCGLFVGCAAILMFKFELNILEL1LQQRCKVTVAFNVVPFVIYAIAKSSEBETYKVLSRITRVKCSAAPMLGKELEDA : 333

At4CL1 : VNAKENAKLLQGGYGMTEAGENVGVAMSLGCAKEFEPFVKSQAGTGVRRNAEMKIVDPDTGDSLSNQPGECICIPGSIKMGYLNNEATAET : 430
At4CL2 : ISAKENAKLLQGGYGMTEAGENVGVAMSLGCAKEFEPFVKSQAGTGVRRNAEMKIVDPDTGDSLSNQPGECICIPGSIKMGYLNNEATAET : 423
At4CL3 : IRRRILCAIQLLQGGYGMTEAGENVGVAMSLGCAKEFEPFVKSQAGTGVRRNAEMKIVDPDTGDSLSNQPGECICIPGSIKMGYLNNEATAET : 433
Lp4CL1 : IRRRILCAIQLLQGGYGMTEAGENVGVAMSLGCAKEFEPFVKSQAGTGVRRNAEMKIVDPDTGDSLSNQPGECICIPGSIKMGYLNNEATAET : 433
Lp4CL2 : IRRRILCAIQLLQGGYGMTEAGENVGVAMSLGCAKEFEPFVKSQAGTGVRRNAEMKIVDPDTGDSLSNQPGECICIPGSIKMGYLNNEATAET : 433
Lp4CL3 : IRRRILCAIQLLQGGYGMTEAGENVGVAMSLGCAKEFEPFVKSQAGTGVRRNAEMKIVDPDTGDSLSNQPGECICIPGSIKMGYLNNEATAET : 433
Os4CL1 : IRRRILCAIQLLQGGYGMTEAGENVGVAMSLGCAKEFEPFVKSQAGTGVRRNAEMKIVDPDTGDSLSNQPGECICIPGSIKMGYLNNEATAET : 406
Zm4CL1 : IRRRILCAIQLLQGGYGMTEAGENVGVAMSLGCAKEFEPFVKSQAGTGVRRNAEMKIVDPDTGDSLSNQPGECICIPGSIKMGYLNNEATAET : 406

Primer-R
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 switchgrass

What I did ... Condontree for looking into codon bias

[Website link]
http://bioweb.pasteur.fr/seqanal/interfaces/codontree.html
Before my starting to clone genes ...

<table>
<thead>
<tr>
<th>Am. Acid</th>
<th>Codon</th>
<th>Number</th>
<th>Freq. %</th>
<th>Cod-Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leu</td>
<td>CTA</td>
<td>5</td>
<td>0.37</td>
<td>0.05</td>
</tr>
<tr>
<td>Leu</td>
<td>CTG</td>
<td>42</td>
<td>3.12</td>
<td>0.39</td>
</tr>
<tr>
<td>Leu</td>
<td>CTT</td>
<td>43</td>
<td>3.20</td>
<td>0.40</td>
</tr>
<tr>
<td>Leu</td>
<td>TTA</td>
<td>5</td>
<td>0.37</td>
<td>0.05</td>
</tr>
<tr>
<td>Leu</td>
<td>TTG</td>
<td>5</td>
<td>0.37</td>
<td>0.05</td>
</tr>
<tr>
<td>Ser</td>
<td>AGC</td>
<td>19</td>
<td>1.41</td>
<td>0.18</td>
</tr>
<tr>
<td>Ser</td>
<td>AGT</td>
<td>4</td>
<td>0.30</td>
<td>0.04</td>
</tr>
<tr>
<td>Ser</td>
<td>TCA</td>
<td>16</td>
<td>1.19</td>
<td>0.16</td>
</tr>
<tr>
<td>Ser</td>
<td>TCT</td>
<td>28</td>
<td>2.08</td>
<td>0.27</td>
</tr>
<tr>
<td>Ser</td>
<td>TCC</td>
<td>19</td>
<td>1.41</td>
<td>0.18</td>
</tr>
<tr>
<td>Ser</td>
<td>TCT</td>
<td>17</td>
<td>1.26</td>
<td>0.17</td>
</tr>
<tr>
<td>Arg</td>
<td>AGA</td>
<td>14</td>
<td>1.04</td>
<td>0.17</td>
</tr>
<tr>
<td>Arg</td>
<td>AGG</td>
<td>17</td>
<td>1.26</td>
<td>0.21</td>
</tr>
<tr>
<td>Arg</td>
<td>CGA</td>
<td>4</td>
<td>0.30</td>
<td>0.06</td>
</tr>
<tr>
<td>Arg</td>
<td>CCG</td>
<td>19</td>
<td>1.41</td>
<td>0.23</td>
</tr>
<tr>
<td>Arg</td>
<td>CGG</td>
<td>22</td>
<td>1.64</td>
<td>0.27</td>
</tr>
<tr>
<td>Arg</td>
<td>CGT</td>
<td>6</td>
<td>0.45</td>
<td>0.07</td>
</tr>
<tr>
<td>Gly</td>
<td>GGA</td>
<td>11</td>
<td>0.82</td>
<td>0.11</td>
</tr>
<tr>
<td>Gly</td>
<td>GGS</td>
<td>50</td>
<td>3.72</td>
<td>0.52</td>
</tr>
<tr>
<td>Gly</td>
<td>GGG</td>
<td>25</td>
<td>1.86</td>
<td>0.26</td>
</tr>
<tr>
<td>Gly</td>
<td>GGT</td>
<td>10</td>
<td>0.74</td>
<td>0.10</td>
</tr>
<tr>
<td>Val</td>
<td>GTA</td>
<td>6</td>
<td>0.45</td>
<td>0.05</td>
</tr>
<tr>
<td>Val</td>
<td>GTT</td>
<td>38</td>
<td>2.83</td>
<td>0.32</td>
</tr>
<tr>
<td>Val</td>
<td>GTG</td>
<td>58</td>
<td>4.31</td>
<td>0.50</td>
</tr>
<tr>
<td>Val</td>
<td>GTT</td>
<td>15</td>
<td>1.12</td>
<td>0.13</td>
</tr>
</tbody>
</table>

...
Before my starting to clone genes ...

4CL gene cloning from switchgrass

primer-F: 5’-CACCACCCGGCCTGCGGAAGGGCGT-3’

primer-R: 5’-CGGGCGAGATCTGGATCAGGGG-3’
Bioinformatics and Its Applications in Plant Biology

Seung Yoo Rhee, Julie Dickerson, and Dong Xu

1 Department of Plant Biology, Carnegie Institution, Stanford, California 94305; email: yose@stanford.edu
2 Plant Center for Computational Biology, Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011-006; email: jules@iastate.edu
3 Horticulture Laboratory, Computer Science Department and Life Sciences Center, University of Missouri-Columbia, Columbia, Missouri 65211-0450; email: mdxu@missouri.edu

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.
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!