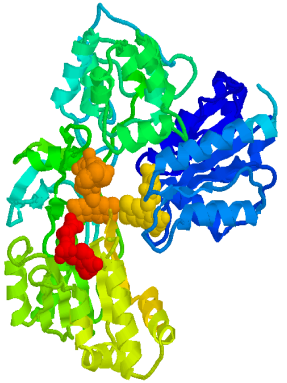
Genome-scale **Protein Function Prediction** using Phylogenomics, Data Integration and Lexical Scoring, *applied on the genomes of* tomato (*Solanum lycopersicum*) and the leguminous plant *Medicago truncatula*

by
Asis Hallab
Max Planck Institute for Plant Breeding Research and
INRES (Institut für Nutzpflanzenwissenschaft und Resourcenschutz, Universität Bonn)

MAX-PLANCK-GESELLSCHAFT

INRES
Institut für
Nutzpflanzenwissenschaften
und Ressourcenschutz

Introduction: Functional Annotation of predicted Proteins



```
>bgh05332_mRNA_polypeptide
MRPSPRHLAKGFLGRSVDEFKRLSTSSMFTGPKKILVTVNLTILLVLKAEGLREPTKPYI
LASFRDSKAIQDCKAMSDKDIGGFSTANLDWVPPSKNQTPNATGSSHGHAKFHGNISIEL
PINRPEVHRTGYAAWRTKDKGYTIFGKTLWDIDPYEFLALRIKSDGRKYFINLQTESIVP
TDIHQHRLYAKRPGEWETLFVPWTEFVRTNHGVVVEPQREMLRQSLRTIGIGLTDRVPGN
FELCIERMWATNEMKNDDSGFE*
>bgh05347_mRNA_polypeptide
MQPLNPFLKAFFKSALPAQCTPVQNHVSSAINALKARVFLQLQVLLVPTTEVFFTSHDSE
```

PhyloFun → 

GO:0000278 mitotic cell cycle

AHRD → 

> G2/mitotic-specific cyclin

Introduction: Functional Annotation of predicted Proteins

Integrated Methods

1. Motif / Pattern Search for conserved Domains

InterPro

Protein Archive

Introduction: Functional Annotation of predicted Proteins

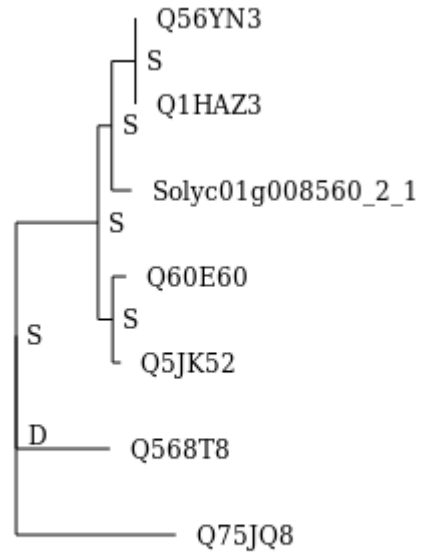
Integrated Methods

2. Sequence Similarity Search in selected Databases

```
Query= bghP002743000001001 STI1 Heat Shock Protein STI1
         (573 letters)

Database: sprot_batches_min_1000_2011_min_solyc.fasta
           530,347 sequences; 188,015,589 total letters

Searching..................................................done



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

sp|Q9USI5|STI1_SCHPO Heat shock protein sti1 homolog OS=Schizosa...   566   e-160
sp|P15705|STI1_YEAST Heat shock protein STI1 OS=Saccharomyces ce...   525   e-148
sp|Q3ZBZ8|STIP1_BOVIN Stress-induced-phosphoprotein 1 OS=Bos tau...   398   e-110
sp|O35814|STIP1_RAT Stress-induced-phosphoprotein 1 OS=Rattus no...   397   e-109
```

Introduction: Functional Annotation of predicted Proteins

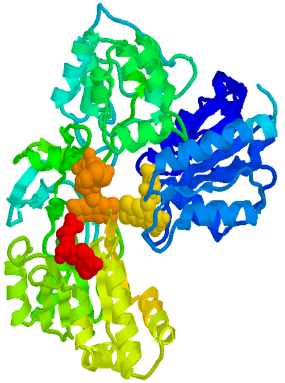
Integrated Methods
3. Phylogenetic Reconstruction



D : Duplication
S : Speciation

PHYLIP

1.Pipeline: (AHRD)
Automated Assignment of Human Readable Descriptions



```
>bgh05332_mRNA_polypeptide
MRPSPRHLAKGFLGRSVDEFKRLSTSSMFTGPKKILVTVNLTILLVLKAEGLREPTKPYI
LASFRDSKAIQDCKAMSDKDIGGFSTANLDWVPPSKNQTPNATGSSHGHAKFHGNISIEL
PINRPEVHRTGYAAWRTKDKGYTIFGKTLWDIDPYEFLALRIKSDGRKYFINLQTESIVP
TDIHQHRLYAKRPGEWETLFVPWTEFVRTNHGVVVEPQREMLRQSLRTIGIGLTDRVPGN
FELCIERMWATNEMKNDDSGFE*
>bgh05347_mRNA_polypeptide
MQPLNPFLKAFFKSALPAQCTPVQNHVSSAINALKARVFLQLQVLLVPTTEVFFTSHDSE
```

AHRD

> G2/mitotic-specific cyclin

1.Pipeline: AHRD

- Interleukin-1 receptor-associated kinase-like 2
- Adenylyl-sulfate kinase; AltName: Full=APS kinase
- Protein FAM190A
- 6,7-dimethyl-8-ribityllumazine synthase
- UPF0059 membrane protein BCAH187_A5502
- UPF0059 membrane protein BCE33L5024
- UPF0059 membrane protein BT9727_5008
- Zinc finger protein 598
- Probable serine/threonine-protein kinase DDB_G0293276
- Vacuolar membrane-associated protein IML1
- Uncharacterized protein C1orf198

1.Pipeline: AHRD

- Interleukin-1 receptor-associated kinase-like 2
- Adenylyl-sulfate kinase; AltName: Full=APS kinase
- Protein FAM190A
- 6,7-dimethyl-8-ribityllumazine synthase
- UPF0059 membrane protein BCAH187_A5502
- UPF0059 membrane protein BCE33L5024
- UPF0059 membrane protein BT9727_5008
- Zinc finger protein 598
- Probable serine/threonine-protein kinase DDB_G0293276
- Vacuolar membrane-associated protein IML1
- Uncharacterized protein C1orf198

1. Pipeline AHRD

NCBI BLAST

Swissprot

TAIR

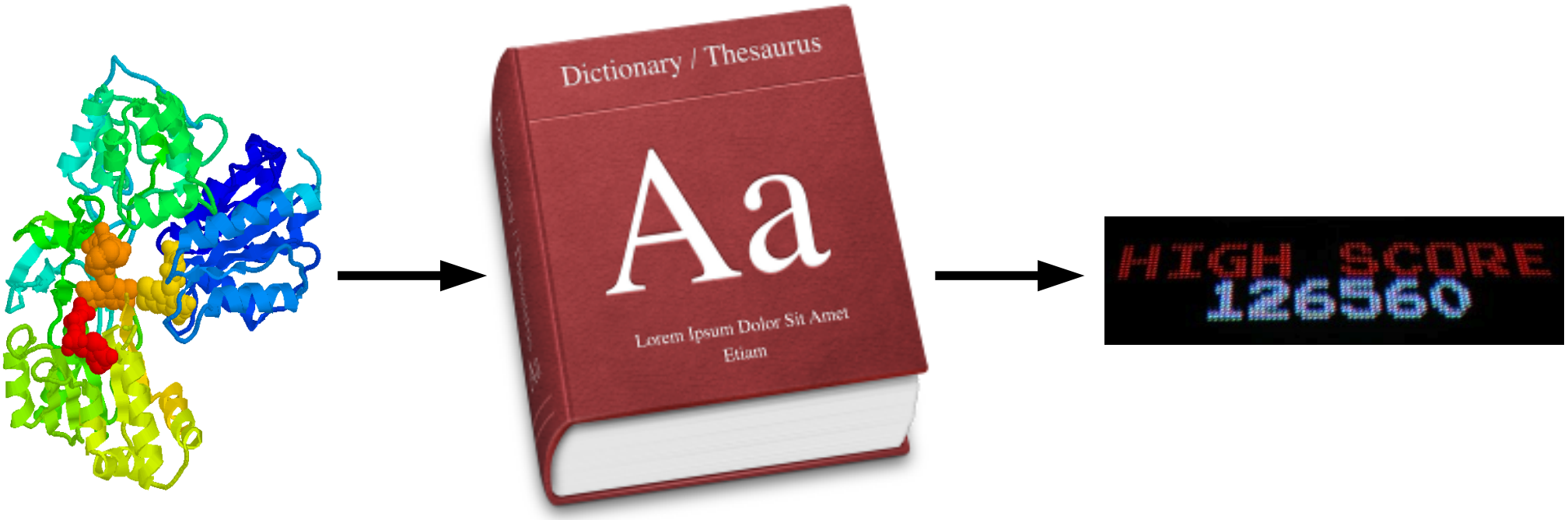
trEMBL

&

GO:0000278 mitotic cell cycle

InterPro
Protein Archive

1. Pipeline AHRD



AHRD's lexical approach:

- Dictionary of meaningful words
- Score words on
  - Frequency,
  - Alignment-quality,
  - Trust put into it's source,
  - Appearance in Gene-Ontology
- Description scored based on contained words

1. Pipeline AHRD

[EXPERTS]	2-dehydropantoate 2-reductase
[AHRD]	2-dehydropantoate 2-reductase
[Blast2GO]	uncharacterized protein
[Swissprot]	Meiotically up-regulated gene 72 protein
[TAIR]	-
[trEMBL]	Putative uncharacterized protein

1. Pipeline AHRD

[EXPERTS] Protein kinase

[AHRD] Protein kinase

[Blast2GO] uncharacterized protein

[Swissprot] CTD kinase subunit alpha

[TAIR] (Cyclin-dependent kinase C;1); kinase

[trEMBL] Putative uncharacterized protein

1. Pipeline AHRD

1400
Blumeria graminis
expert annotated
proteins

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}$$

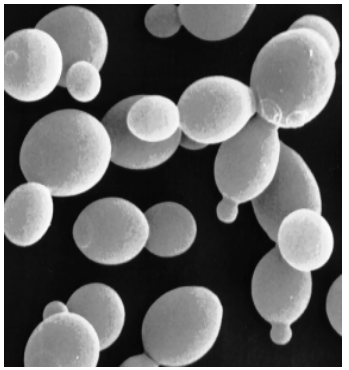
AHRD performed better than

best

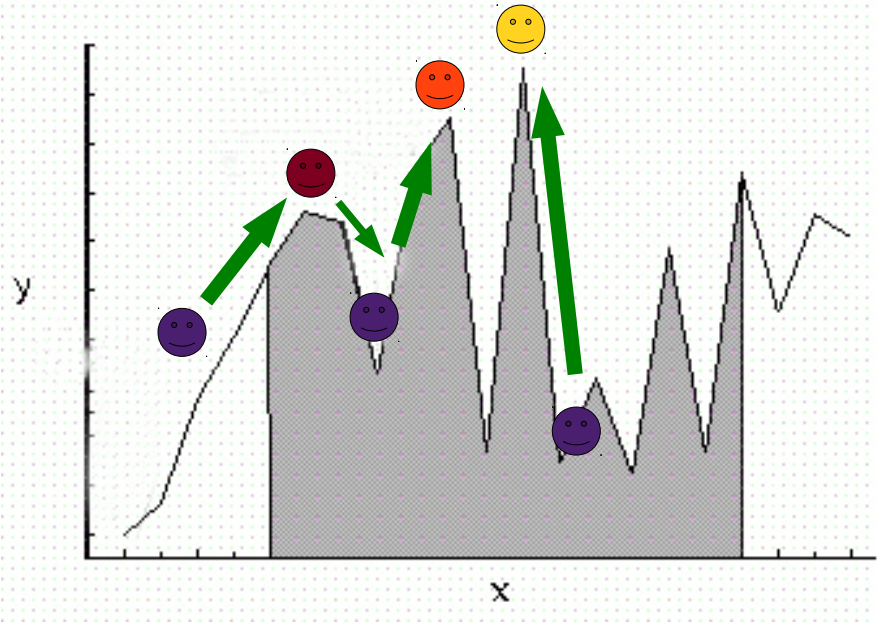
- Swissprot
- trEMBL

1. Pipeline AHRD

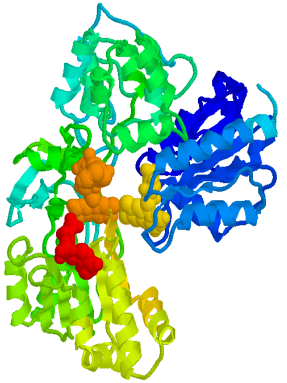
Generic usage of Blast-Databases:



Parameter-Optimization using simulated annealing:



2.Pipeline: PhyloFun



```
>bgh05332_mRNA_polypeptide
MRPSPRHLAKGFLGRSVDEFKRLSTSSMFTGPKKILVTVNLTILLVLKAEGLREPTKPYI
LASFRDSKAIQDCKAMSDKDIGGFSTANLDWVPPSKNQTPNATGSSHGHAKFHGNISIEL
PINRPEVHRTGYAAWRTKDKGYTIFGKTLWDIDPYEFLALRIKSDGRKYFINLQTESIVP
TDIHQHRLYAKRPGEWETLFVPWTEFVRTNHGVVVEPQREMLRQSLRTIGIGLTDRVPGN
FELCIERMWATNEMKNDDSGFE*
>bgh05347_mRNA_polypeptide
MQPLNPFLKAFFKSALPAQCTPVQNHVSSAINALKARVFLQLQVLLVPTTEVFFTSHDSE
```

PhyloFun

GO:0000278 mitotic cell cycle

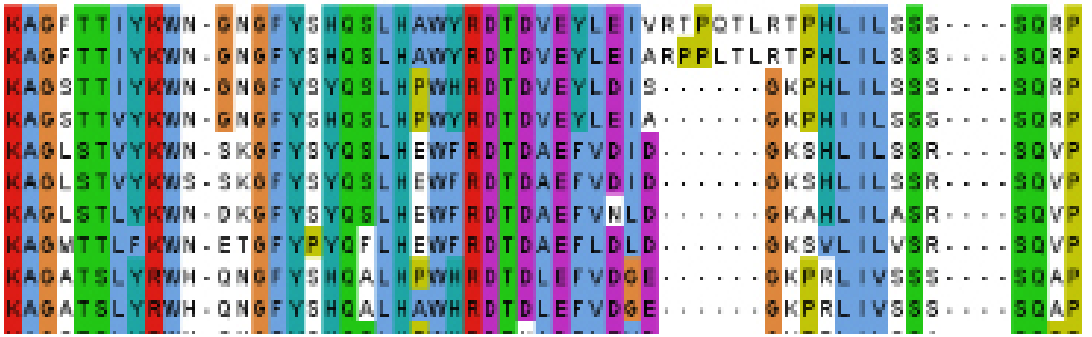
2.Pipeline: PhyloFun

1. Blast-Search
2. Multiple Sequence Alignment
3. Filter for conserved positions
4. Reconstruct Phylogeny
5. Find Speciation / Duplication-Events
6. Assign candidate functions

```
Query= bghP002743000001001 STI1 Heat Shock Protein STI1
         (573 letters)

Database: sprot_batches_min_1000_2011_min_solyc.fasta
           530,347 sequences; 188,015,589 total letters

Searching.................................................done



                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

sp|Q9USI5|STI1_SCHPO Heat shock protein sti1 homolog OS=Schizosa...   566   e-160
sp|P15705|STI1_YEAST Heat shock protein STI1 OS=Saccharomyces ce...   525   e-148
sp|Q3ZBZ8|STIP1_BOVIN Stress-induced-phosphoprotein 1 OS=Bos tau...   398   e-110
sp|O35814|STIP1_RAT Stress-induced-phosphoprotein 1 OS=Rattus no...   397   e-109
```

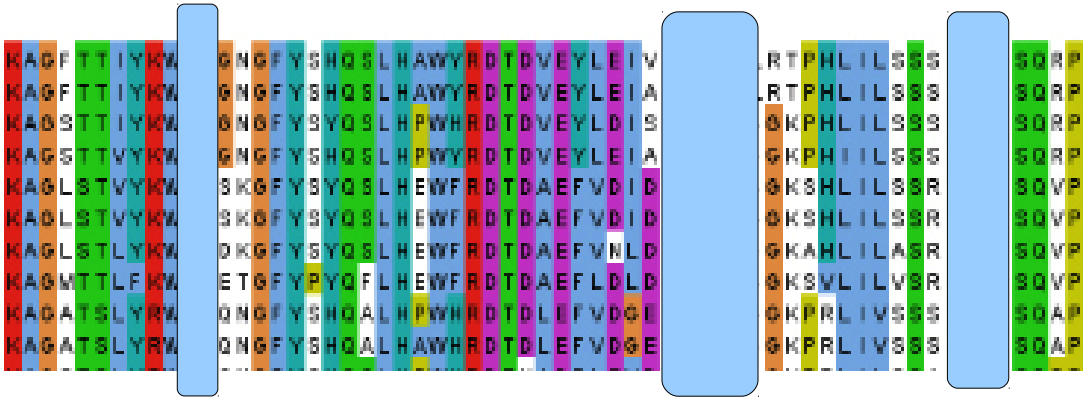
2.Pipeline: PhyloFun

1. Blast-Search
2. Multiple Sequence Alignment
3. Filter for conserved positions
4. Reconstruct Phylogeny
5. Find Speciation / Duplication-Events
6. Assign candidate functions



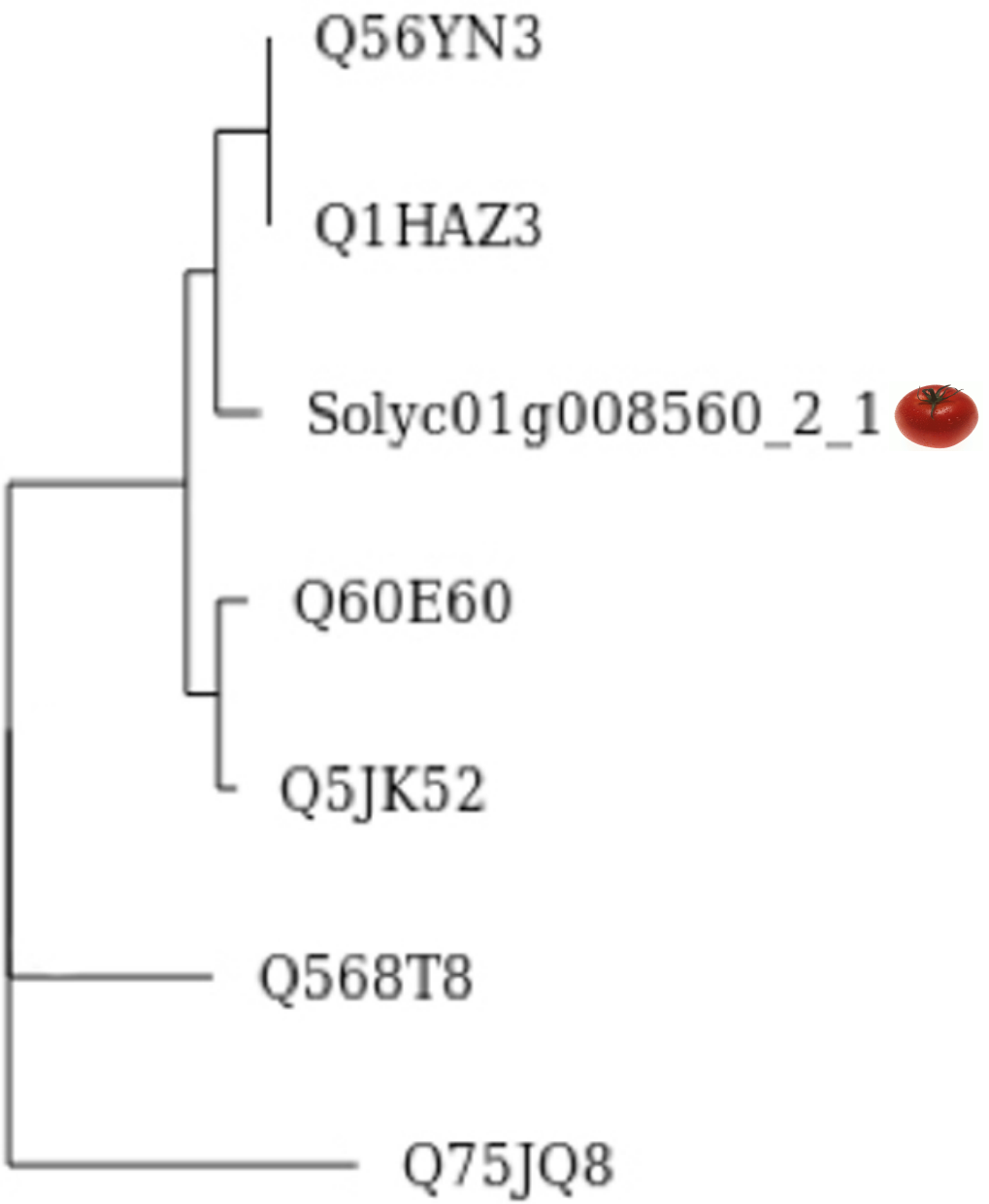
2.Pipeline: PhyloFun

1. Blast-Search
2. Multiple
   Sequence Alignment
3. Filter for
   conserved positions
4. Reconstruct Phylogeny
5. Find Speciation /
   Duplication-Events
6. Assign candidate
   functions



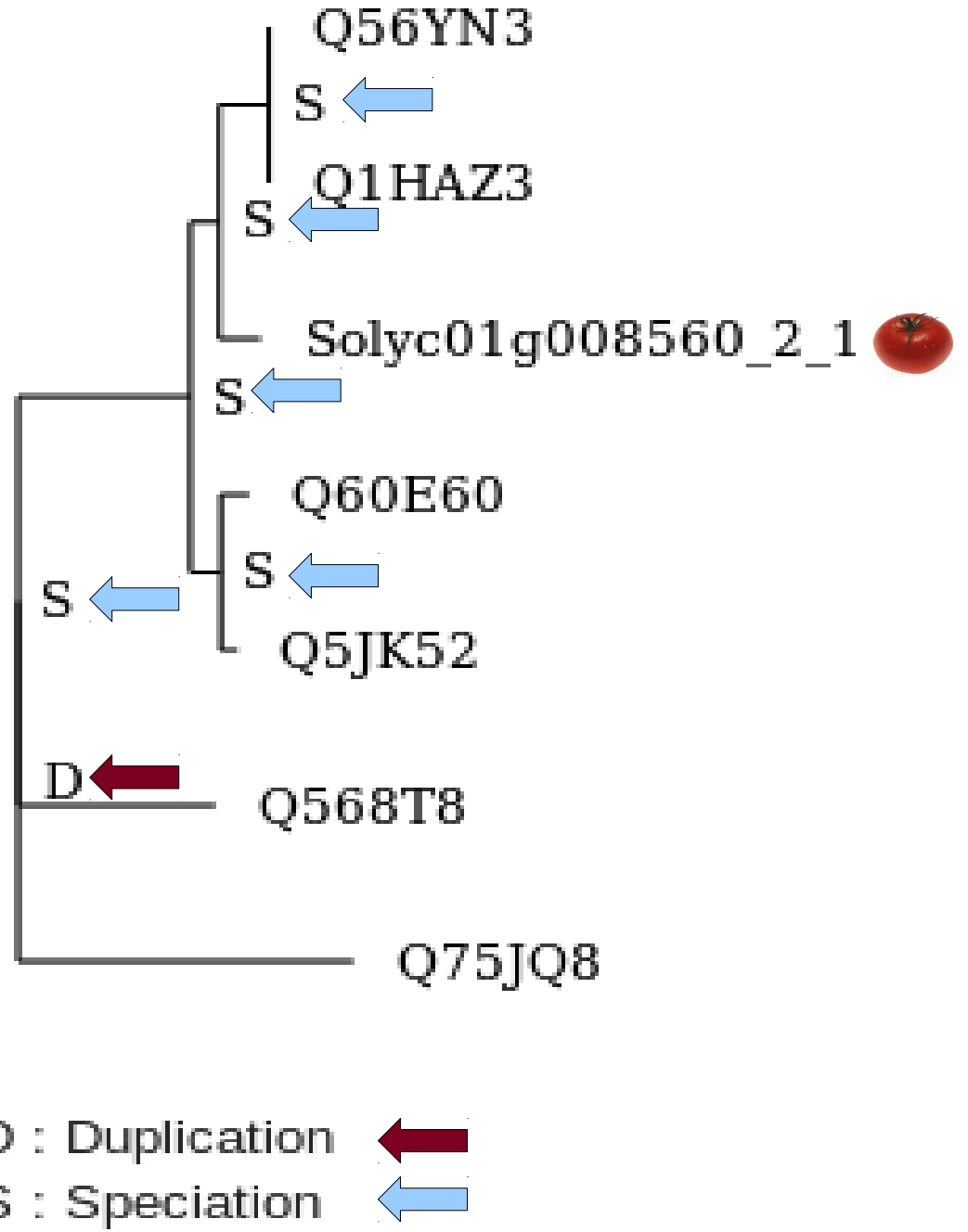
2.Pipeline: PhyloFun

1. Blast-Search
2. Multiple Sequence Alignment
3. Filter for conserved positions
4. Reconstruct Phylogeny
5. Find Speciation / Duplication-Events
6. Assign candidate functions



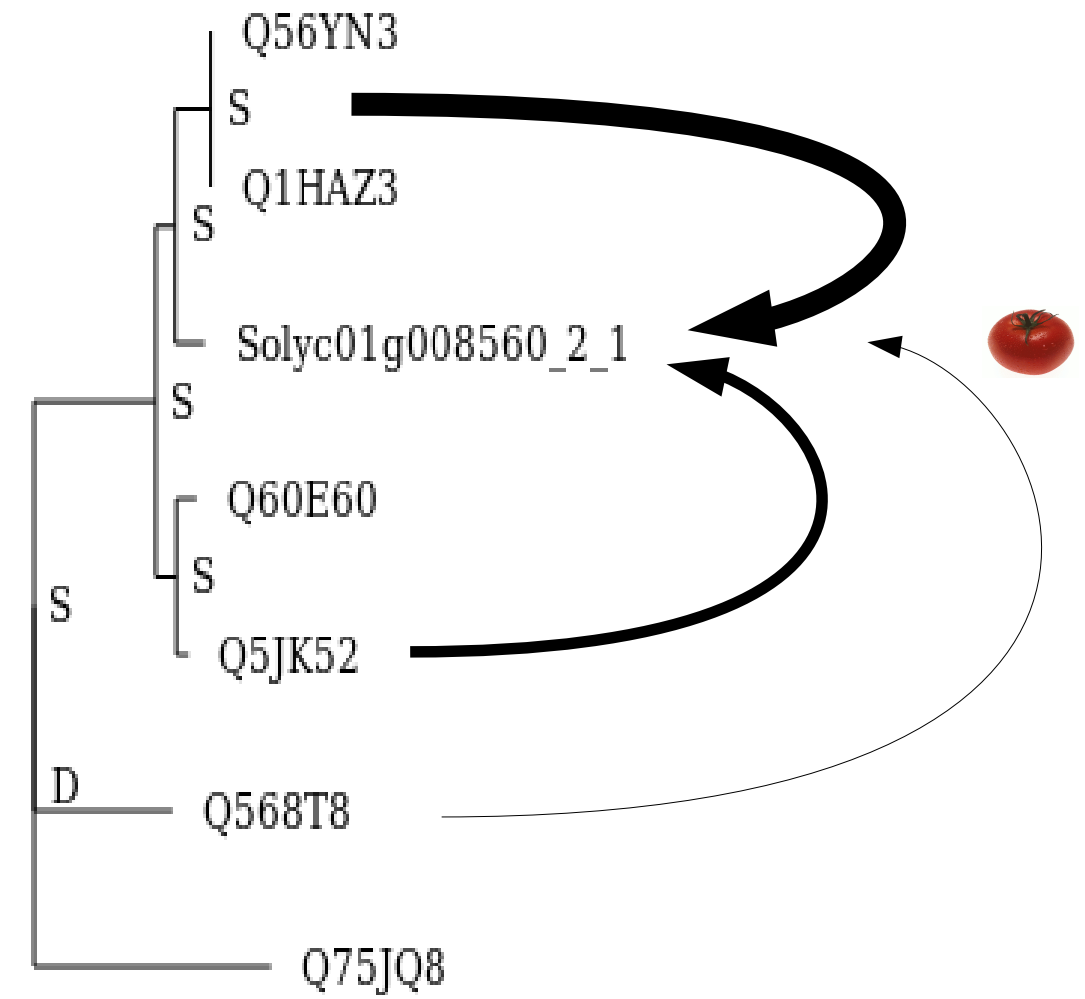
2.Pipeline: PhyloFun

1. Blast-Search
2. Multiple Sequence Alignment
3. Filter for conserved positions
4. Reconstruct Phylogeny
5. Find Speciation / Duplication-Events
6. Assign candidate functions



2.Pipeline: PhyloFun

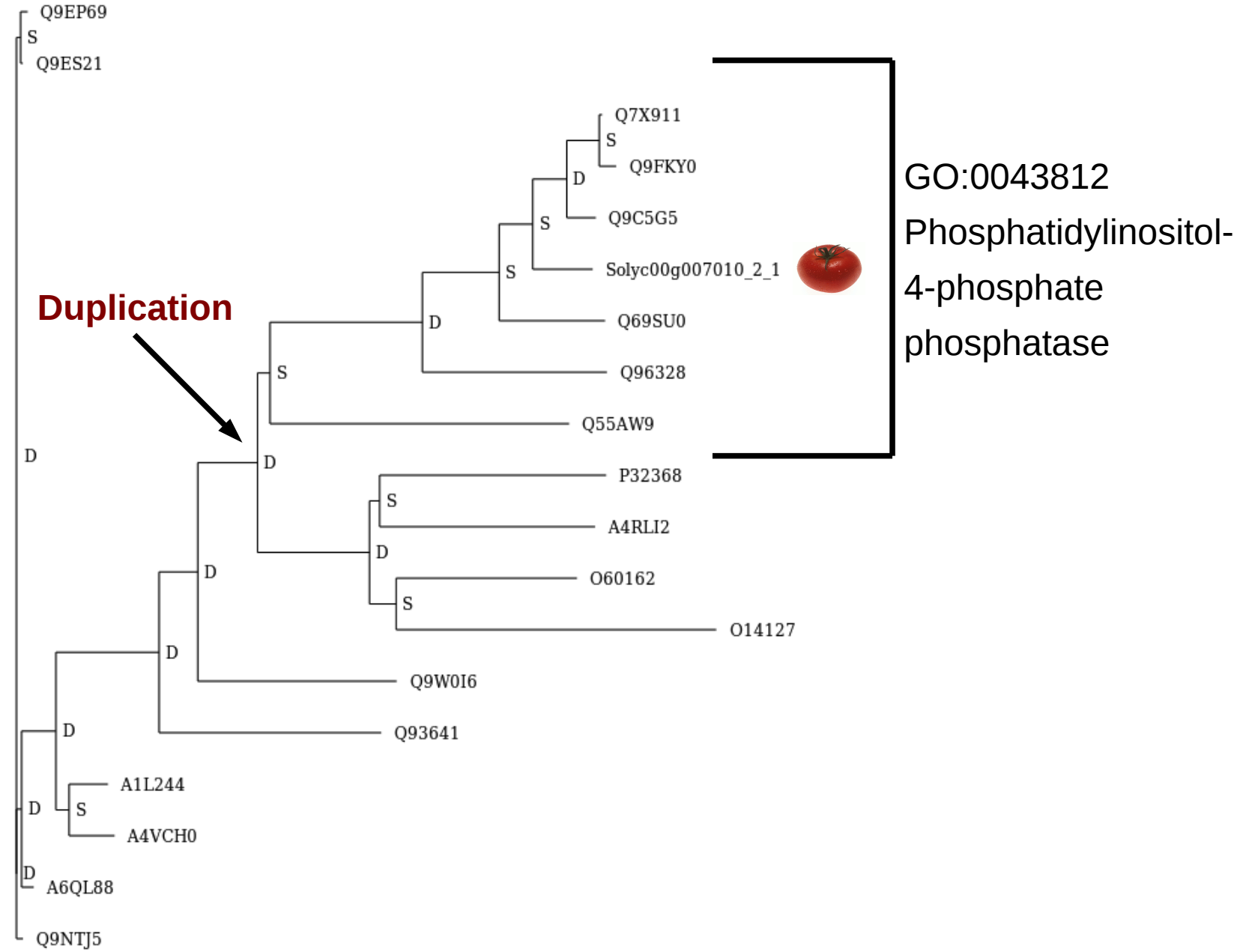
1. Blast-Search
2. Multiple Sequence Alignment
3. Filter for conserved positions
4. Reconstruct Phylogeny
5. Find Speciation / Duplication-Events
6. Assign candidate functions



D : Duplication
S : Speciation

SIFTER - Engelhardt et al. 2005 PLoS Computational Biology 1(5):e45

Phosphatase-Protein-Family



D : Duplication
S : Speciation

All annotations available at our website

Solyc00g007010% Search

Search for organism, accession or namespace.
Use '%' as wildcard.

Menu

Protein Overview
GO Annotations
InterPro Annotations
RPS-BLAST Annotations
BLAST Hits

BLAST Hits for: Solyc00g007010.2.1 (itag)

Phosphatidylinositol phosphate (PtdInsP) phosphatase involved in hydrolysis of PtdIns (AHRD V1 **** [...] *Solanum lycopersicum*

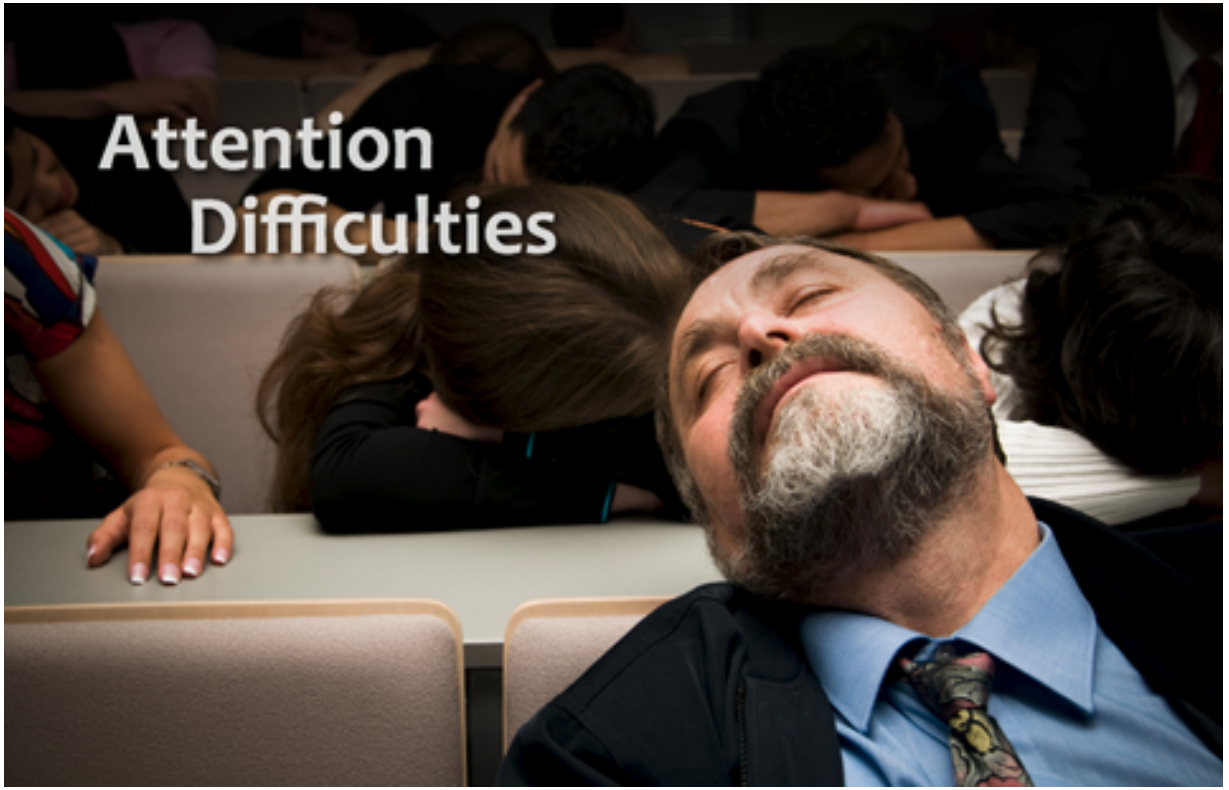
- [ ] Overlap >= 70%
- [x] BLAST Hits share InterPro domain with protein
- [ ] BLAST Hits with verified molecular function
- [ ] BLAST Hits with verified biological process
- [ ] BLAST Hits with verified cellular component

Filter

Hit protein	Description	Organism	Bit score / Evalue	Identity / Overlap	GO term(s)	InterPro domain(s)
B9H663 (uniprot)	B9H663_POPTR Predicted protein OS=Populus trichocarpa GN=POPTRDRAFT_558773 PE=4 SV=1	*Populus trichocarpa*	945.0 / 0.0	75% / 99%	GO:0042578 (phosphoric ester hydrolase activity)	IPR002013 (Synaptojanin, N-terminal)
D7U476 (uniprot)	D7U476_VITVI Whole genome shotgun sequence of line PN40024, scaffold_44.assembly12x (Fragment) OS=Vi[...]	*Vitis vinifera*	945.0 / 0.0	79% / 100%	GO:0042578 (phosphoric ester hydrolase activity)	IPR002013 (Synaptojanin, N-terminal)
D3Y5N9 (uniprot)	D3Y5N9_BRACM SAC-like protein OS=Brassica campestris GN=BrSAC1 PE=2 SV=1	*Brassica campestris*	940.0 / 0.0	73% / 99%	GO:0042578 (phosphoric ester hydrolase activity)	IPR002013 (Synaptojanin, N-terminal)

afawe.mpipz.mpg.de

Thank you!



Heiko Schoof
Kathrin Klee
Sri Girish Srinivasa Murthy
Jens Warfsmann
Haili Song
Anika Jöcker
Andreas Jöcker
The EU-SOL project
International Tomato Annotation Group
International Medicago Genome Annotation Group

MAX-PLANCK-GESELLSCHAFT

eusol

INRES
Institut für
Nutzpflanzenwissenschaften
und Ressourcenschutz