

Functional Annotation

Functional vs Structural Annotation

- Structure:
 - Location of exons/introns/UTRs
 - Identification of genomic elements by position
- Function:
 - What does it do?
 - Biological information

Outline

- Lots of tools!
- Dammit:
 - BLAST to Uniref90
 - HMMER search of Pfam-A
 - Infernal search of Rfam
- InterProScan
- Gene Ontology

dammit!

- Wrapper for:
 1. Transdecoder – translate to amino acids
 2. BLAST to OrthoDB – eukaryotic ortholog database
 - BLAST to BUSCO – Benchmarking Universal Single-Copy Orthologs
 3. BLAST to Uniref90 *
 4. HMMer search of Pfam-A *
 5. Infernal search of Rfam *

Uniref90



- Clustered sets of sequences from the giant database of all known proteins (UniProt Knowledgebase)
- UniRef100 – collapse identical sequences and sub-fragments with 11 or more residues from any organism into a single entry
- UniRef90 – use CD-HIT to collapse sequences that have at least 90% sequence identity to and 80% overlap with the longest sequence

83,050,155



43,405,259

SwissProt



- A curated set of the UniRef database
- 551,705 entries (July 6th, 2016)
- Manual annotation:
 - Identification of homologs w/ BLAST
 - Protein domain id and protein family classification
 - Association with relevant literature
 - Extensive cataloging of information from laboratory experiments
 - Gene Ontology term assignment

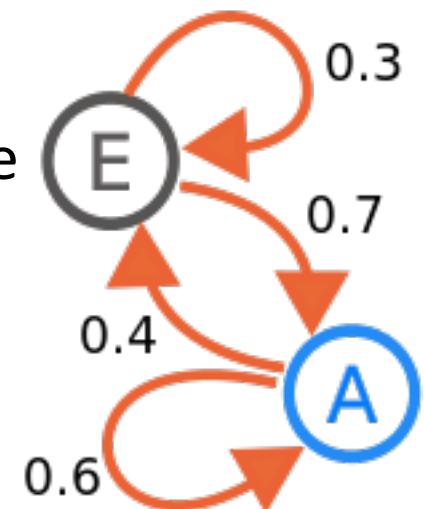
Pfam and HMMER

Pfam + HMMER

- There is more to the world than just BLAST (ie traditional sequence alignment)
- The second most popular algorithm is HMMER.
- HMM = Hidden Markov Model
- But to understand that we need to talk about...

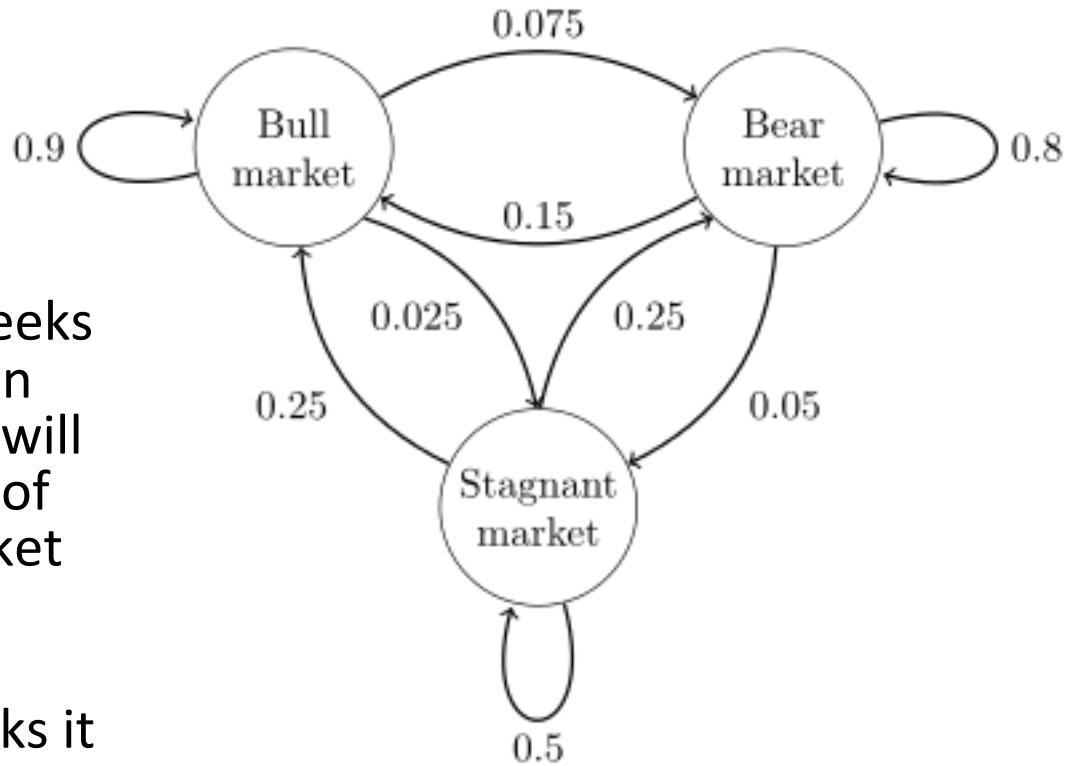
Markov Chain

- A Markov chain is a random process that undergoes transitions from one state to another on a state space
- Has the property of “memorylessness”
- the probability distribution of the next state depends only on the current state and not on the sequence of events that preceded it
- Called the Markov property
- A Markov chain is a type of Markov Model that is fully observable – we know all the states and probabilities for moving between states



Markov Chain

- How is it used statistically?
- Possible to calculate:
- the long-term fraction of weeks during which the market is in each state (62.5% of weeks will be in a bull market, 31.25% of weeks will be in a bear market and 6.25% of weeks will be stagnant)
- the average number of weeks it will take to go from a stagnant to a bull market

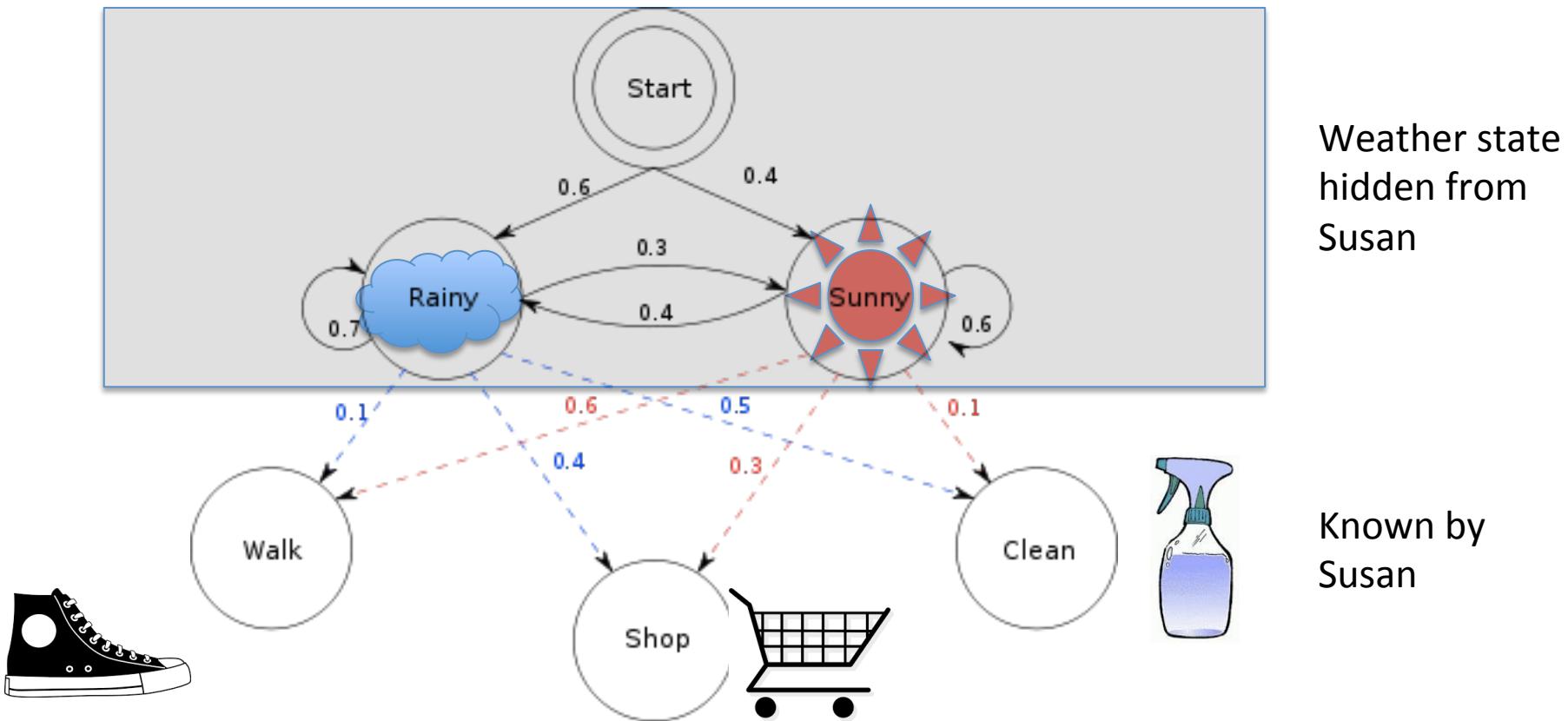


Hidden Markov Model

- The markov chain is only one type of markov model. Another is the hidden Markov model.
- Similar to a Markov chain
- Hidden (unobservable) states
- Example

Hidden Markov Model

- Bob has a friend Susan. Everyday he posts on Facebook weather he is walking, shopping or cleaning. Susan is a mathematician and recognizes this as an HMM.



What does this have to do with biology?

- Allow you to incorporate heterogenous types of information for a problem
 - Allow you to add new information more easily.
 - Gene finding. We should account for:
 - splice-site consensus
 - codon bias
 - exon/ intron length preferences
 - open reading frame analysis
 - HMMs provide a conceptual toolkit for building complex models.
- 
- How should the parameters be set?
How do we weight them?
How to score?
How confident that an answer is correct?

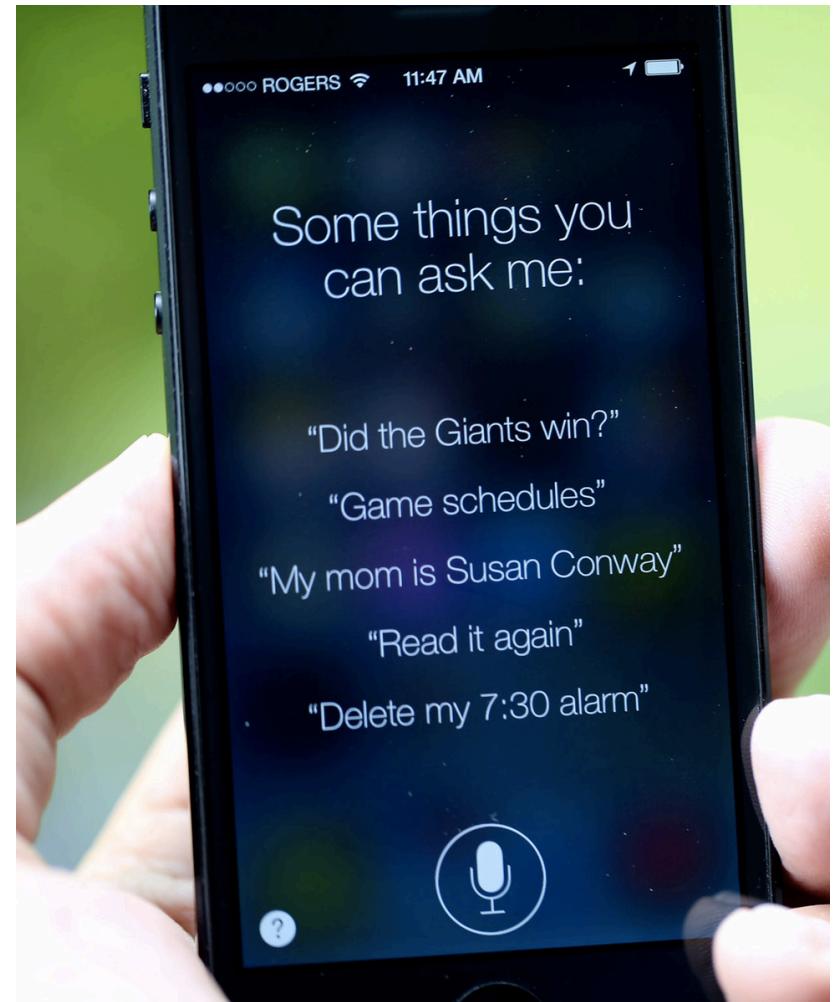
What does this have to do with biology?

Problems often addressed with HMMs:

- Finding a gene
- Searching for a sequence profile
- Multiple sequence alignment
- Regulatory site identification

Outside of biology, best known for temporal pattern recognition:

- Speech
- Handwriting
- Gesture



HMM – 5' splice example

- We have a sequence.

Definitely Exon



CTTAGATCGAAATTGATTTCGTAAAACGTTCCCCGG

?????????

Definitely Intron

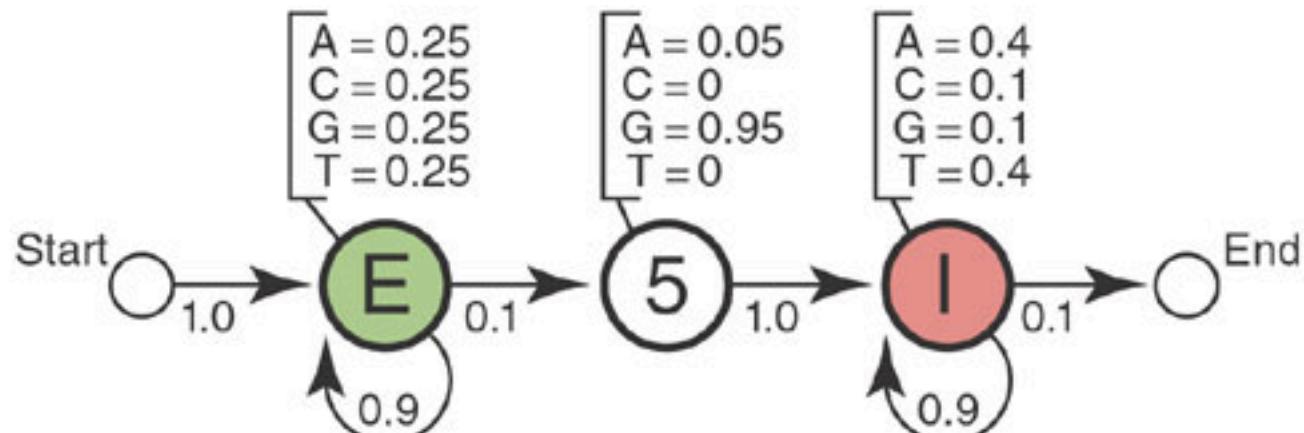


Where is the splice site?

HMM – 5' splice example

- Lets say we know some information about splicing that will be helpful
- exons have a uniform base composition on average (25% each base), while introns are A/T rich (say, 40% each for A/T, 10% each for C/G),
- the 5'SS consensus nucleotide is almost always a G (say, 95% G and 5% A).
- We can make an HMM.
- We have hidden states: each base is an Exon(E), an Intron(I) or a 5'SS(S)
- We need to find the most likely state that produced the observed sequence

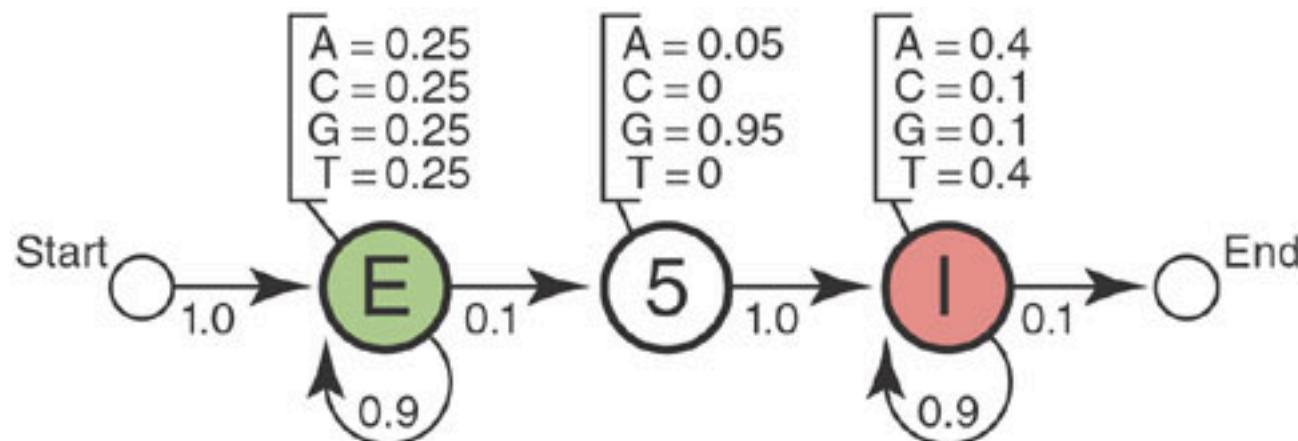
HMM – 5' splice example



Sequence: C T T C A T G T G A A A G C A G A C G T A A G T C A

Lets test different underlying states to see which is the most likely.

HMM – 5' splice example



Sequence: C T T C A T G T G A A A G C A G A C G T A A G T C A

State path: E 5 I I I I I I I log P
-41.22

Parsing:
Eddy
What is a
hidden
markov
model?
2004

-43.90
-43.45
-43.94
-42.58
-41.71



- Start with a multiple sequence alignment
- Feed into **hmmbuild**
 - Generate an **hmm profile**
- Calibrate the model with **hmmc_calibrate**
 - Increase sensitivity of searching
- Search for new homologs that belong to your group with **hmmsearch**



- Why not use BLAST?
- Has much more power in the case of many sequences from the same family – can build a more accurate model of that family by using information about:
 - how conserved each column of the alignment is
 - which residues are most likely at each position
- With a well described protein family, can detect much more remote evolutionary relationships than BLAST.
- Used to be much slower, with new HMMER3 implementation, now is almost as fast as BLAST
- What sorts of databases can we search with HMMER?



- Within a database of protein sequences, many are members of existing protein families and have similar functions. How to organize this information?
- Need to identify protein clusters and to produce multiple sequence alignments.
- The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).
- Originally published in 1997
- Pfam-A = manually curated family data
- Pfam-B = computationally generated family data



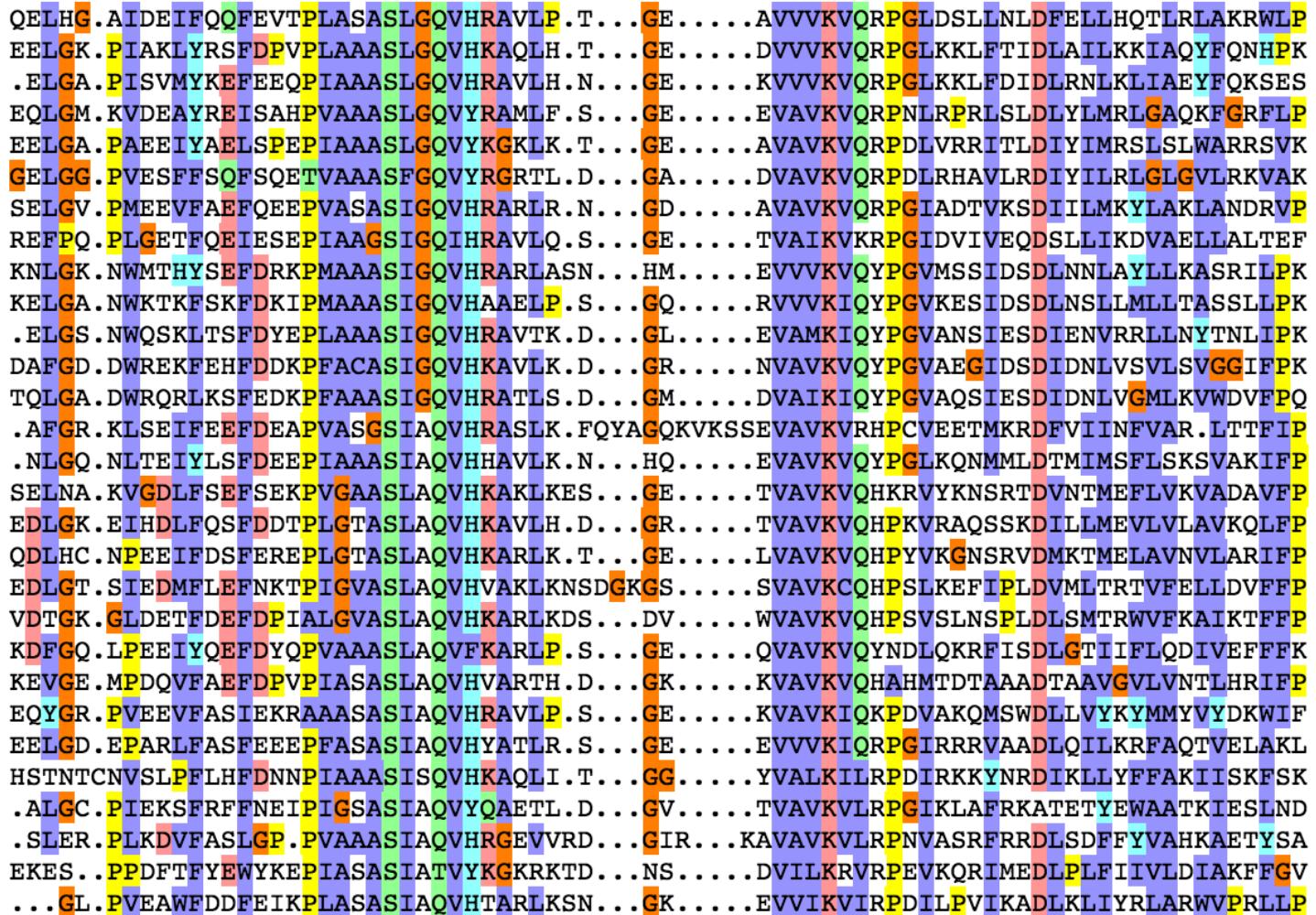
- Currently has 16,306 families (version 30)
- Families are grouped into “clans” - related by similarity of sequence, structure or profile-HMM
- Family information includes gene architecture, structure, sequences from hundreds to thousands of species and interactions.



- HMMER is used by Pfam for two purposes:
 - To construct Pfam clusters
 - To detect matches from a given sequence to a cluster
- The states of the Pfam HMM correlate to the multiple sequence alignment: match, delete, insert
- The matching amino acid probabilities from the HMM can be visualized as a logo

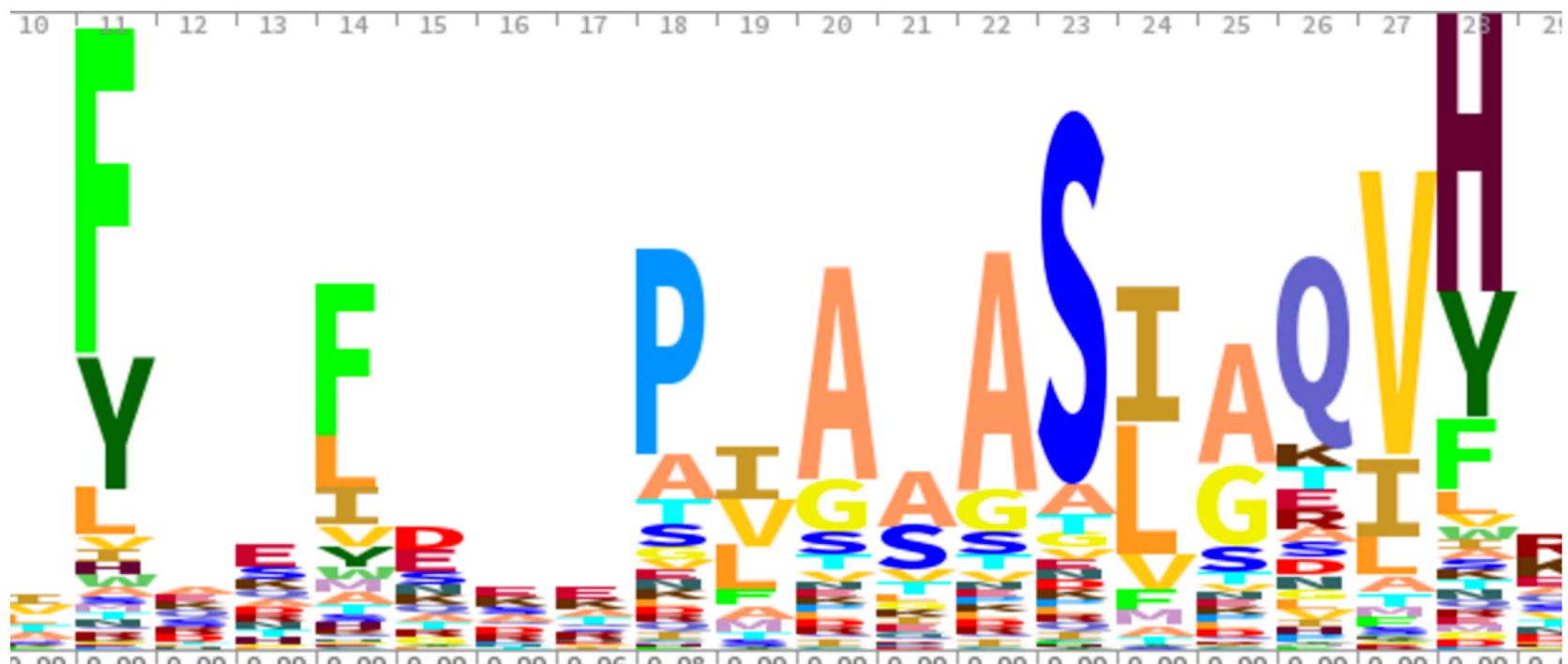
Multiple Sequence Alignment

Y095_SYN3/115-238
Y1770_SYN3/142-261
B9DGY1_ARATH/248-365
Y1919_SYN3/127-246
Y005_SYN3/161-279
O80962_ARATH/256-373
O27682_METTH/119-238
Y889_SYN3/100-218
ABCI_SCHPO/284-401
COQ8_YEAST/176-292
Q9SBB2_ARATH/284-398
COQ8_CAEEL/417-533
Q9VYI6_DROME/336-452
Q3ECK9_ARATH/268-392
F4ID59_ARATH/156-271
O17735_CAEEL/142-259
ADCK1_HUMAN/143-259
Q9W133_DROME/137-253
MCP2_YEAST/166-289
MCP2L_SCHPO/168-286
Q9VTG5_DROME/162-278
Y2090_ARATH/147-268
YF9E_SCHPO/167-287
Y647_MYCTU/150-271
Q9ZCP5_RICPR/34-153
H2VFS0_ZYMMO/111-228
Q89WD1_BRADU/112-231
Y445_PBCV1/90-208
UBIB_SHIDS/115-232



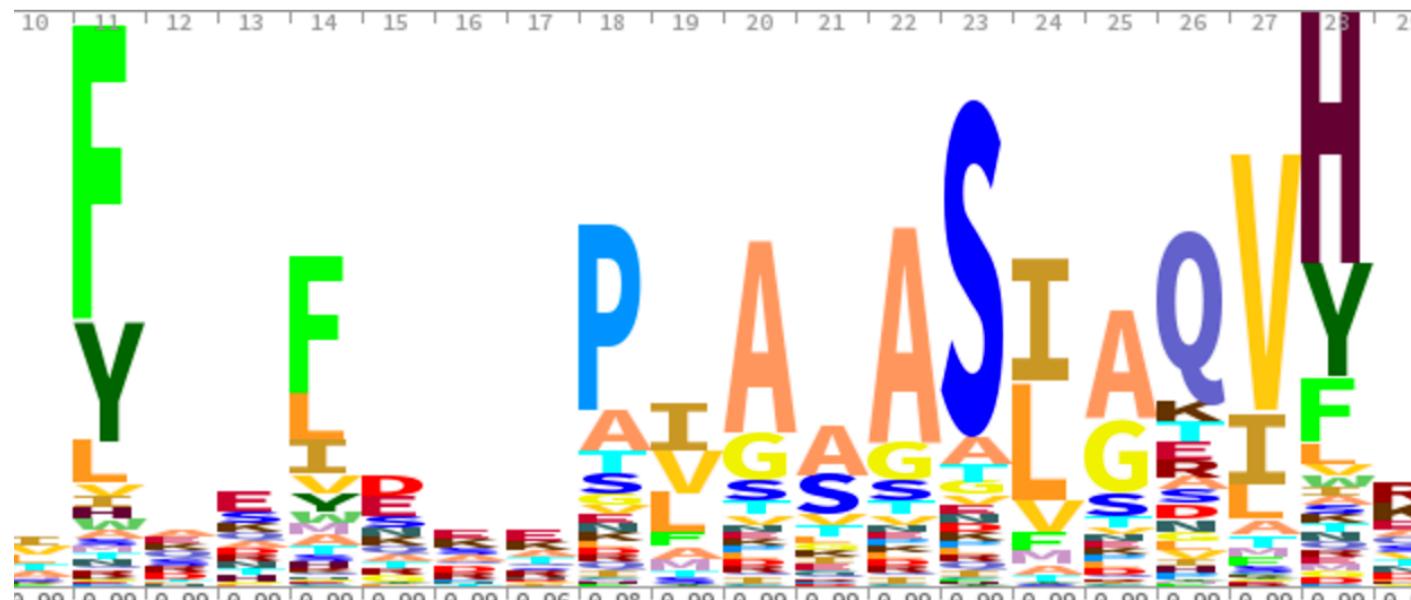
Becomes a Profile

- Represented as a logo



Pfam

- Take all 16,306 profiles
- Use hmmsearch to compare your sequences to these families
- Uses all the logo information!



Example:

- Search green ash protein

ALCLIMLAHSGGGAAISPNSVNTTRPNLPTINDSKQIENSTTPPPTQDQSYSCVCNKAFASYQALGGHKASHRKNATASDDG
NHSTSTTTAAASTASNVSAALNPRGRLHECSICHKSFTGQALGGHKRRHYEGIIGGGSSKSSVTSSDGGASSHAPRDFDLNLPATP
EFQLELTVDVCVKKSQFVGQEVESPMPFKKPR.T.PT.FGERF

- Results

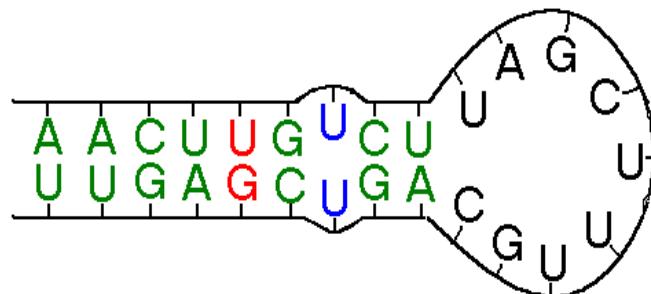
Family	Description
<u>zf-C2H2_6</u>	C2H2-type zinc finger
#HMM	heCdeCsksFpS1qaLggHkksHrk
#MATCH	+ C++C+k F S+qaLggHk+sHrk
#PP	78*****8
#SEQ	YSCSVNCNAFASYQALGGHKASHRK

Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value
		Start	End	Start	End	From	To			
Domain	<u>CL0361</u>	53	79	54	78	2	26	27	45.9	3.2e-12

Infernal + Rfam



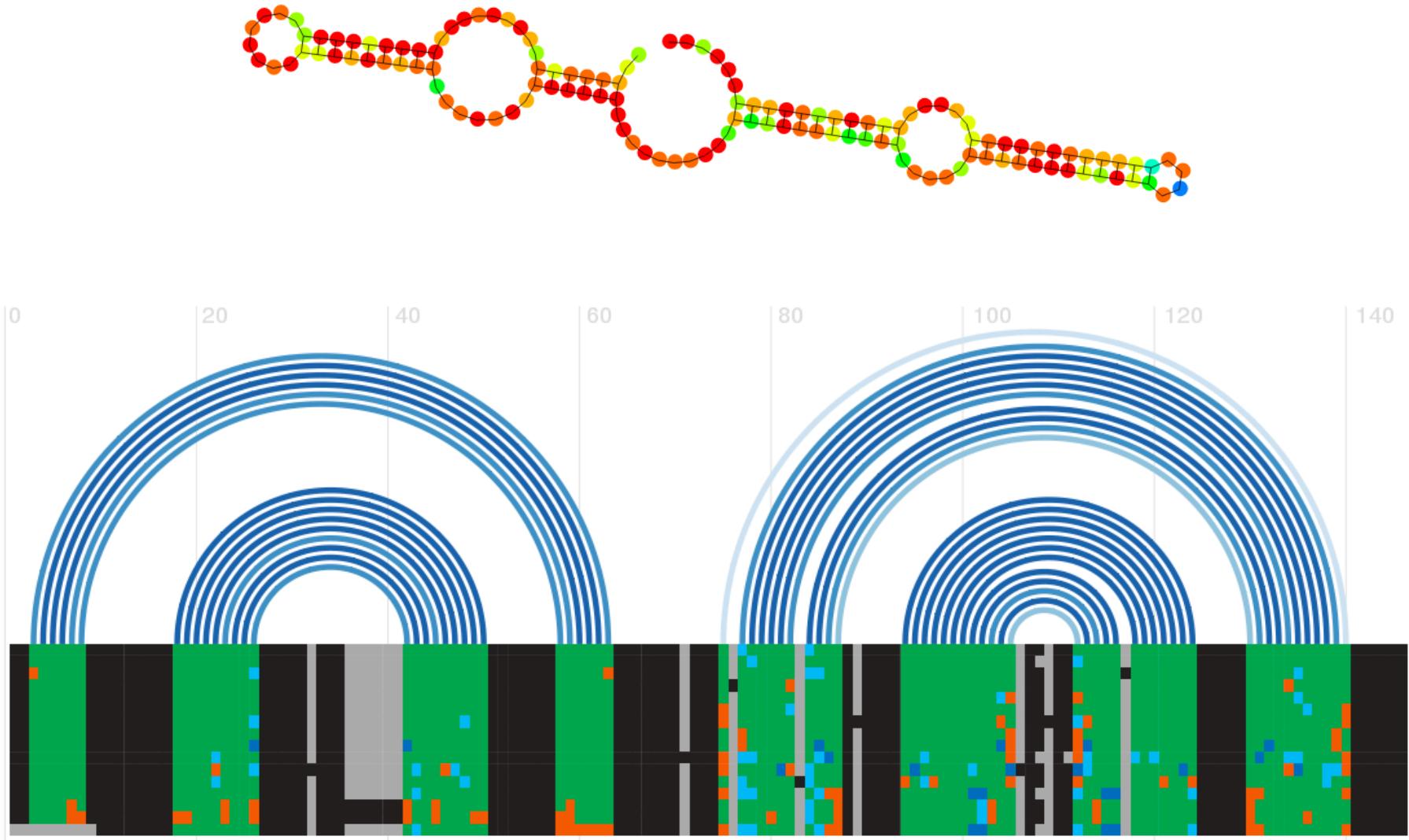
- Infernal ("INFERence of RNA ALignment")
- Tool for searching for RNA structure and sequence similarities
- Uses information about sequence AND structure
- Also uses HMM (generates covariance model instead of a profile)
- Start with:
 - Multiple sequence alignment
 - Special annotation of bases that are paired to create the secondary structure



Example snoRNA

AAAGCAGGUUGCAAUUACAGUGCUCUCAUUU.GUG.....GAAGUACUGCCAUUAUCCUGCUGAAAGAA.AAGC.CGUGUU.AAUCA.UUUUUGAUUUUGCCUU.UA
AAAGCAGGUAGCAAUUACAGUGCUCUCAUUU.GUG.....GGAGUACUGCCAUUAUCCUGCUGAAAGAA.AAGC.CAUGUU.GGUUG.UUUCUGAUUUUGCCUU.U-
-----AGCAAUUACAGUGCUCUCAUUU.GUG.....GGAGUACUGCCAUUAUCCUGCUGAAAGAA.AAGC.CAUGUU.GGUUG.UUUCUGAUUUUGCCUU.U-
AGAGCAGGUUGCAAUUACAGUGCUCUCAUUU.GUG.....GAAGUACUGCCAUUAUCCUGCUGAAAGAA.AAGC.UAUGUU.GAUCA.UUUUUGAUUUUGCCUU.C-
AAAGCAGGUUGCAAUUACAGUGCUCUCAUUU.GUG.....GAAGUACUGCCAUUAUCCUGCUGAAAGAA.AAGC.UGUGUU.GAUCG.UUAUUGAUUUUGCCC.UA
AAAGCAGGUUGCAAUUACAGUGCUCUUCGUUU.GUG.....GAAGUACUGACAUUAUCCUGCUGAAAGAA.AAAC.AGUGUU.GAUCA.UUUUUGAUUUUGCCUC.UC
AAAGCAGGUUGCAAUUACAGUGCUCUUCUUU.GUG.....GAAGUAUUGACAUUAUCCUGCUGAAAGAA.AAUC.UGUGUU.GAUCGuUUUUUGAUUUUGCCAU.UUa
UACGCAGGUUGCAAUUACAGUGCUCUUGUUU.GGG.....GAAGUACUGCUGUUAUCCUGCUGAAAGAC.AAGC.UGUGUU.AGUCA.UUUUUGAUUUUGCCUU.UA
AAAGCAAGCUGCAAUUACAGUGCUCUCAUUU.GUGaaaacUAAAACUGCCAUUAUCCUGCUGAAAGAA.AAGC.UGUGUU.AAUGA.UUUUUGAUUUUGCCUU.UG
AAAGCAGGCUGCAAUUACAGUACUUCAGUUU.GUG.....GAAGUACUGCCAUUAUCCUGCUGAGAGAAAGC.CAUGUU.GGCCG.GCUCUGGUUUUGCCUC.U-
UGAGCAGGUUGCAGUCCAGUCUUUGUUUcGUG.....GGAGUGCUGGCAUAACCCUGCUGAAAACA.AAUA.UGUGCC.AAUCA.UUUUUUUAUUUACCUCaUU.
UAAGCAGGUUGCAAUUACAGUGCUCUCAUUU.GUG.....GAAGUACUGACAUUAUCCUGCUGAAAGAA.AAUCAUGUGUG.GAUCA.UUUUUGAUUUUGCCUU.UG
AAAGCAGGUUGCAAUUACAGUACUUCAUUCU.GUG.....GAAGUAUUGCCAUUAUCCUGCUGAAAGAA.AAGC.CGUGUUuAAUCA.UUUCGGGUUUUGCCUG.UA
AAAGCAGGUUGCAAUUACAGUGCUCUCAUUU.GUG.....GAAGUACUGACAUUAACCCUGCUGAAAGAA.AAUG.UGUGUC.GAUCA.UUUUUGAUUUUGCCUU.UA
AAAGCAGCUGGAAUUGCAGUGCUCUCAUUU.GUGaaaacUAAAACCAUCAUUAUGCUGCUGAAAGAA.AAGC.UGUUUU.AAUGA.UUUUUGAUUUUGCCUU.UG
AAAGCAGGUUGCAAUUACAGUGCUCUUAUUU.GUG.....AAAGUACUGUCAUUAUCCUGCUGAAAGAA.AAGC.UGUGUU.GGUCC.UUUUUGAUUUUGCCAC.UG

Example snoRNA





- RNA types:
 - non-coding RNA genes
 - structured cis-regulatory elements
 - self-splicing RNAs
- Grouped into RNA families, each represented by:
 - multiple sequence alignments
 - consensus secondary structures
 - covariance models (CMs).

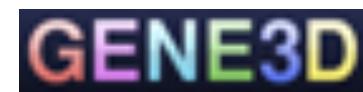
dammit!

- Wrapper for:
 1. Transdecoder – translate to amino acids
 2. BLAST to OrthoDB – eukaryotic ortholog database
 - BLAST to BUSCO – Benchmarking Universal Single-Copy Orthologs
 3. BLAST to Uniref90 *
 4. HMMer search of Pfam-A *
 5. Infernal search of Rfam *

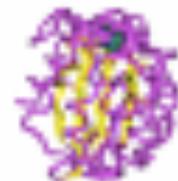
InterProScan

InterPro

- InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites
- uses predictive models, known as signatures, provided by several different databases



Superfamily



Ontology

Ontology

- Roots in philosophy – how we conceptualize and specify knowledge (Aristotle)
 - Super useful to teaching things to computers
 - This is a very big area of thought and utility, we're going to focus on a relatively simple example:
 - Controlled Vocabulary
- Example:
- Wine
 - White Wine
 - Rose Wine
 - Red Wine
 - Beaujolais
 - Red Burgundy
 - Red Zinfandel
 - Merlot

From NCBI SRA for Arabidopsis

I want sequences that relate to flower structures. I have to hand pick:

- Inflorescense
- Inflorescence
- Immature inflorescence
- Flower
- Flowers
- Pistils pollinated for 8 Hours
- 3xHA_inflorescence_biological_replication1
- 3xHA_inflorescence_biological_replication2
- 3xHA-VvCEB1-OX_inflorescence_biological_replication3

Plant Structure Ontology

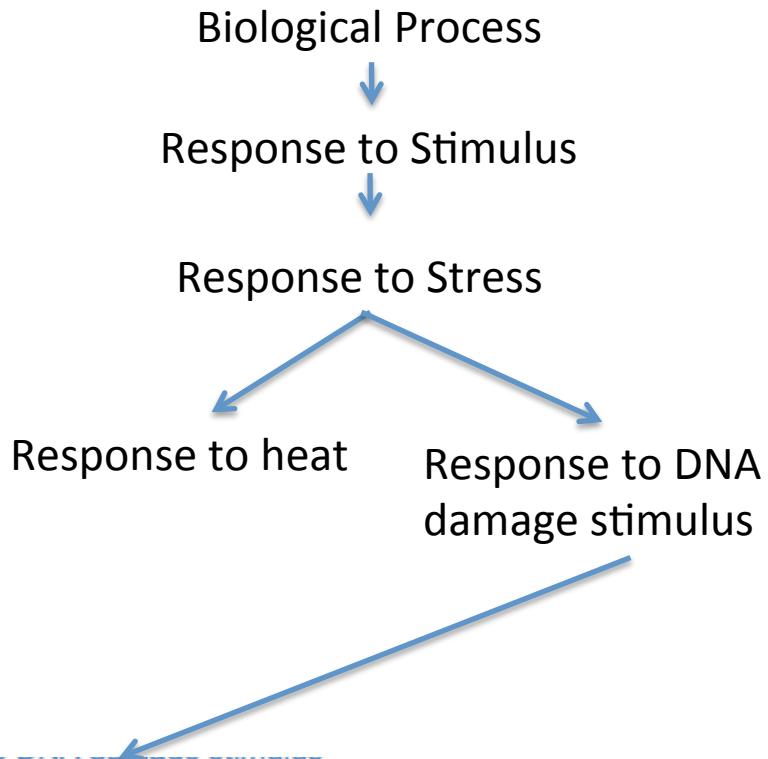
- All these would be coded in a computer readable structure:
 - Inflorescence
 - Flower
 - Gynoecium (Pistil)
 - Androecium
 - Perianth



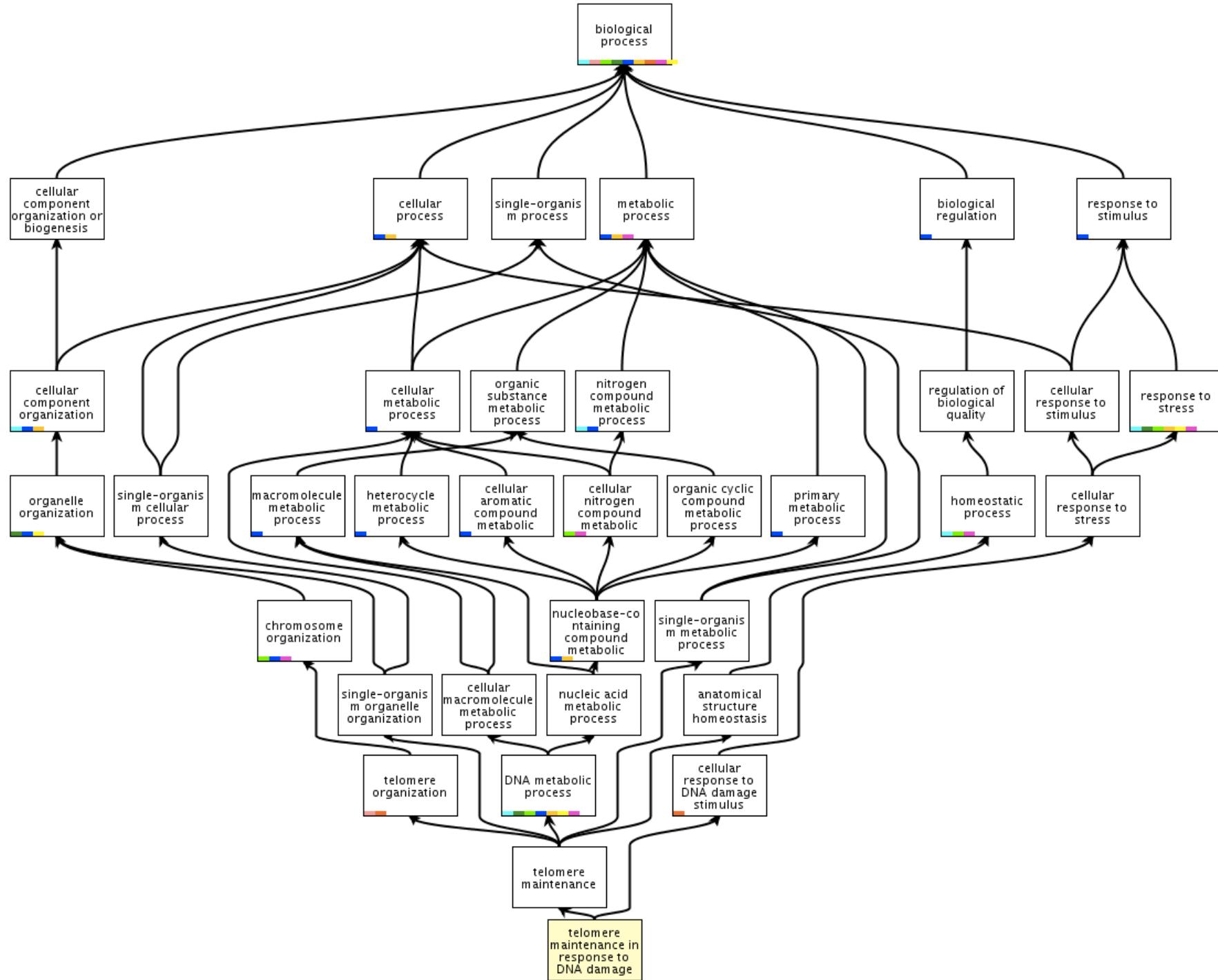
GENEONTOLOGY

Unifying Biology

- Used for annotating genes
- Three sections
 - Biological Processes
 - Metabolic Functions
 - Cellular Components
- Each section is formed as graph or network of terms



GO:2001022 positive regulation of response to DNA damage stimulus
GO:2001020 regulation of response to DNA damage stimulus
GO:1990248 regulation of transcription from RNA polymerase II promoter in response to DNA damage
GO:0031297 replication fork processing
GO:0042770 signal transduction in response to DNA damage
GO:0043247 telomere maintenance in response to DNA damage



Uses of Gene Ontology

- Finding members of the same biological process or pathway
- Finding high level patterns of metabolic or biological activities
- Looking for statistical enrichment of GO terms
 - From the control to the treatment, the occurrence of lignan production related genes increases
- Tools:
 - GO home page
 - BiNGO cytoscape plugin

Example

