

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

July 6th 2016)

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

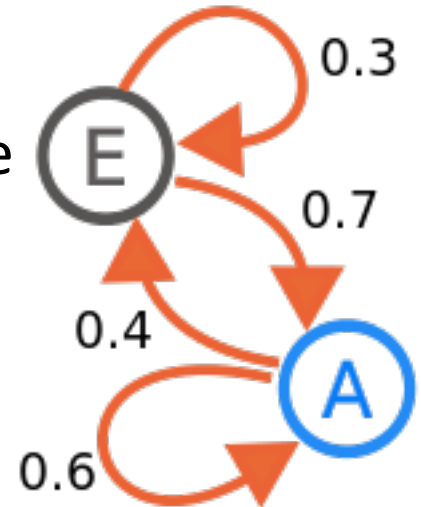
Pfam and HMMER

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...

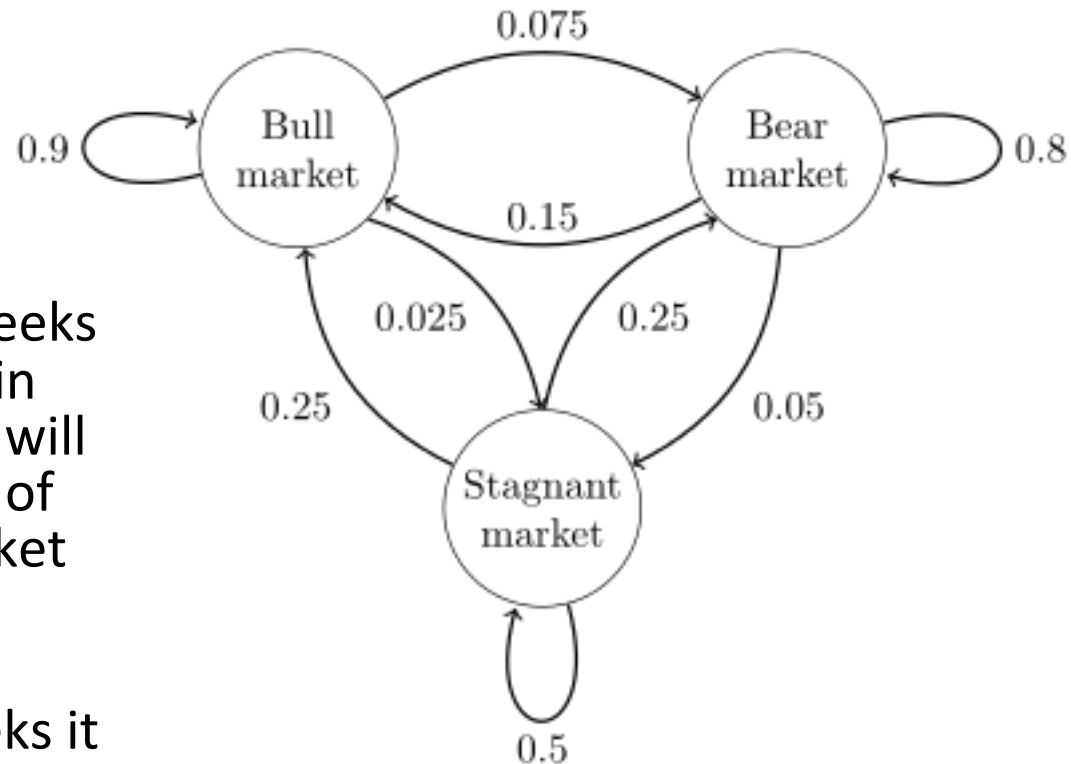
Markov Chain

- 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

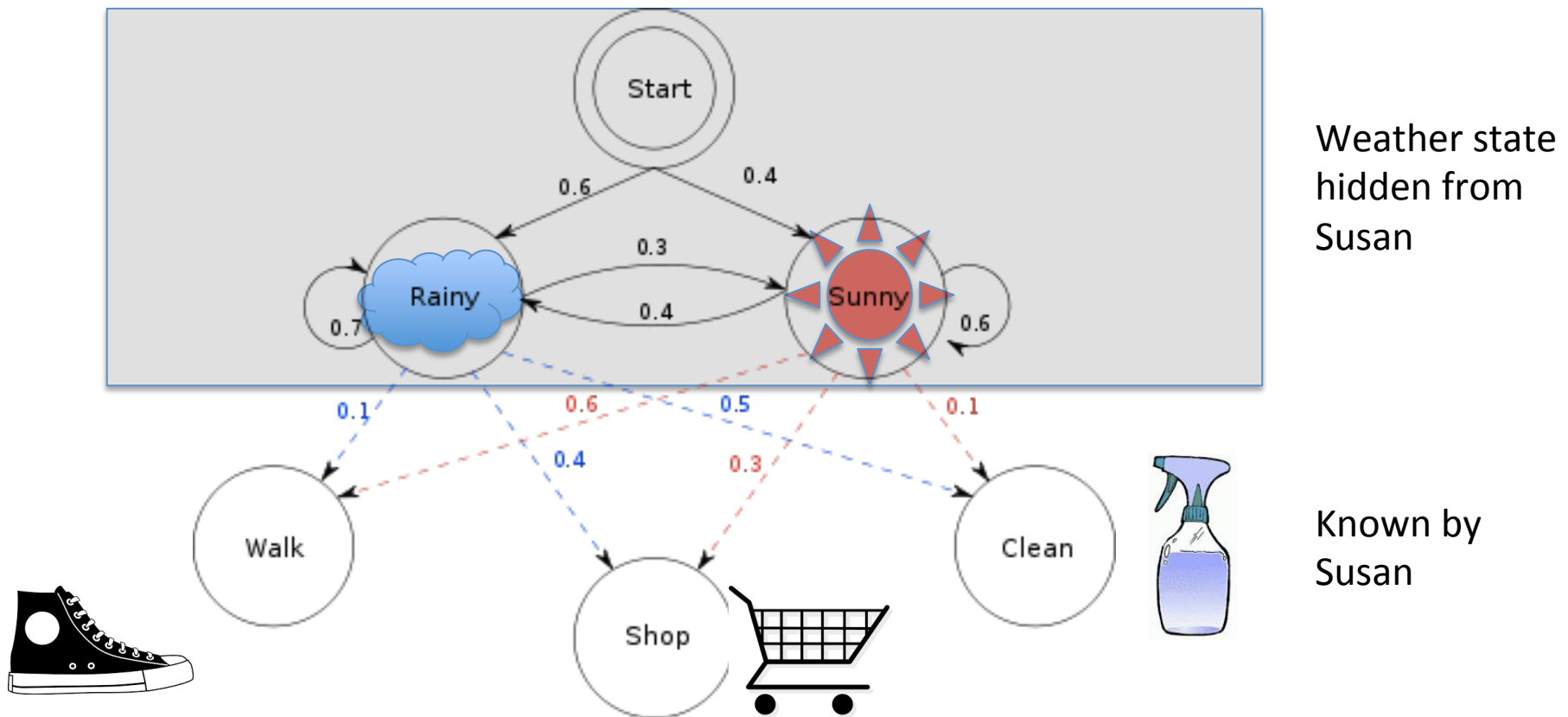


Hidden Markov Model


- 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 heterogeneous 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



Definitely Intron



CTTAGATCGAAATTCGATTTTCGTAAAACGTTCCCCGG

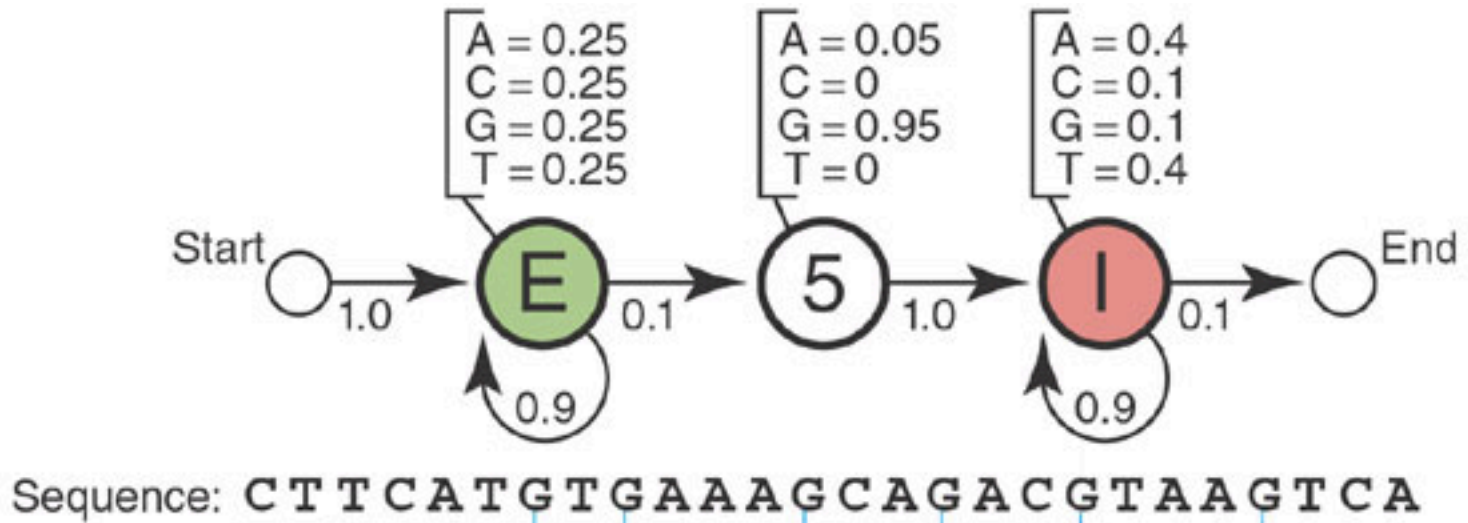
????????

Where is the splice site?

HMM – 5' splice example

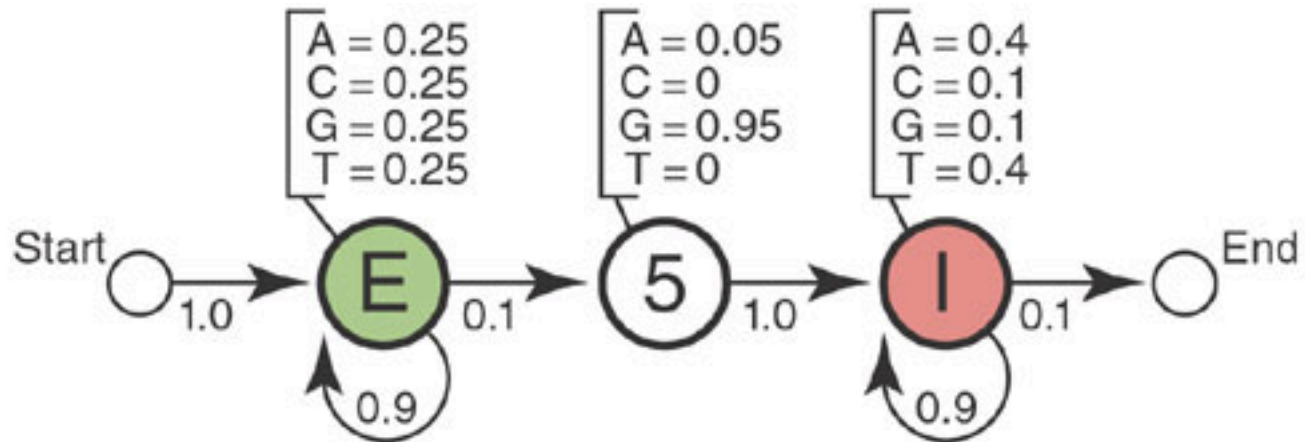
- Lets say we know some information about splicing that will be helpful
- exons have a uniform base composition on average (25% each base), while introns are A/T rich (say, 40% each for A/T, 10% each for C/G),
- the 5'SS consensus nucleotide is almost always a G (say, 95% G and 5% A).
- We can make an 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



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

HMM – 5' splice example



Sequence: **CTTCATGTGAAAGCAGACGTAAGTCA**





- Start with a multiple sequence alignment
- Feed into **hmmbuild**
 - Generate an **hmm profile**
- Calibrate the model with **hmmcalibrate**
 - 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.



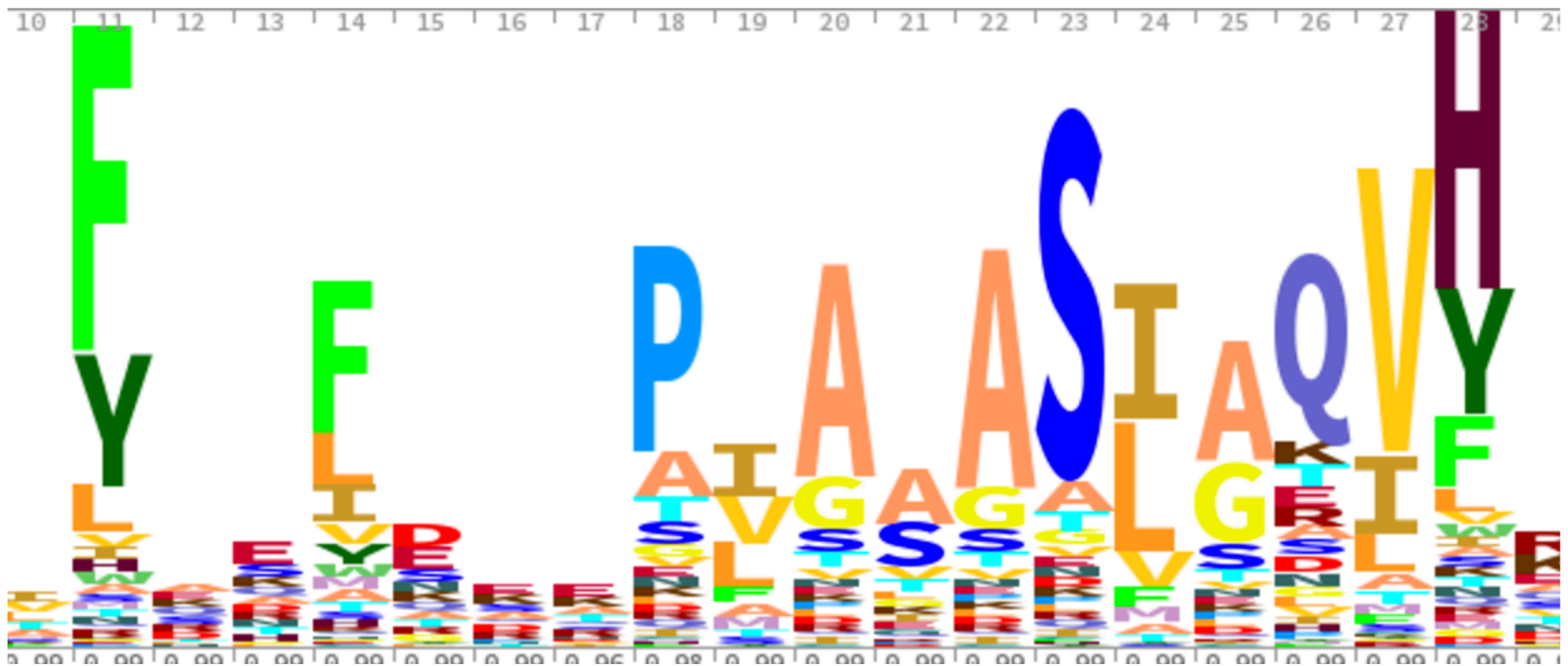
- HMMER is used by Pfam for two purposes:
 - To construct Pfam clusters
 - To detect matches from a given sequence to a cluster
- The states of the Pfam HMM correlate to the multiple sequence alignment: match, delete, insert
- The matching amino acid probabilities from the HMM can be visualized as a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

Multiple Sequence Alignment

```
Y095_SYNY3/115-238      QELHG.AIDEIFQQFEVTPLASASLGQVHRAVLPT...GE.....AVVVKVQRPGLDSLNLNDFELLHQTLRLAKRWLP
Y1770_SYNY3/142-261     EELGK.PIAKLYRSFDPVPLAAASLGQVHKAQLHT...GE.....DVVVKVQRPGLKKLFTIDLAILKKIAQYFQNHFK
B9DGY1_ARATH/248-365    .ELGA.PISVMYKEFEEQPIAAASLGQVHRAVLHN...GE.....KVVVKVQRPGLKKLFDIDLRLNLKLIAYFQKSES
Y1919_SYNY3/127-246     EQLGM.KVDEAYREISAHPVAAASLGQVYRAMLF.S...GE.....EVAVKVQRPNLRPRLSLDLYLMRLGAQKFGRFLP
Y005_SYNY3/161-279      EELGA.PAEIYAELSPEPIAAASLGQVYKKGKLT...GE.....AVAVKVQRPDLVRRITLDIYIMRSLSLWARRSVK
O80962_ARATH/256-373    GELGG.PVESFFSQFSQETVAAASFGQVYRGRTL.D...GA.....DVAVKVQRPDLRHAVLRDIYILRLGLGVLRKVAK
O27682_METTH/119-238    SELGV.PMEEVFAEFQEEPVASASIGQVHRARLR.N...GD.....AVAVKVQRPGIADTVKSDIILMKYLAKLANDRVP
Y889_SYNY3/100-218      REFPO.PLGETFQEIIESEPIAAGSIGQIHRAVLQS...GE.....TVAIKVKRPGIDVIVEQDSLIIKDVALLALTEF
ABCI_SCHPO/284-401      KNLGK.NWMTHYSEFDRKPMAAASIGQVHRARLASN...HM.....EVVVKVQYPGVMSSIDSDLNNLAYLLKASRILPK
COQ8_YEAST/176-292      KELGA.NWKTKFSKFDKIPMAAASIGQVHAAELPS...GQ.....RVVVKIQYPGVKESIDSDLNSLLMLLTASSLLPK
Q9SBB2_ARATH/284-398    .ELGS.NWQSKLTSFDYEPLAAASIGQVHRAVTK.D...GL.....EVAMKIQYPGVANSIESDIENVRRLNNTNLIPK
COQ8_CAEL/417-533       DAFGD.DWREKFEHFDDKPFACASIGQVHKAVLKD...GR.....NVAVKVQYPGVAEGIDSDIDNLVSVLSVGGIFPK
Q9VYI6_DROME/336-452    TQLGA.DWRQRLKSFEDKPFAAASIGQVHRATLS.D...GM.....DVAIKIQYPGVAQSIESIDIDNLVGMKLVWDVFPQ
Q3ECK9_ARATH/268-392    .AFGR.KLSEIFEEFDEAPVASGSIAQVHRASLK.FQYAGQKVKSSSEVAVKVRHPCVEETMKRDFVIINFVAR.LTTFIP
F4ID59_ARATH/156-271    .NLGQ.NLTEIYLSFDEEPIAAASIAQVHHAVLKN...HQ.....EVAVKVQYPGQKQNMMLDTMIMSFLSKSVAKIFP
O17735_CAEL/142-259     SELNA.KVGDLFSEFSEKPVGAASLAQVHKAKLKES...GE.....TVAVKVQHHRVYKNSRTDVNTMEFLVKVADAVFP
ADCK1_HUMAN/143-259     EDLGK.EIHDLFQSFDDTPLGTASLAQVHKAVLHD...GR.....TVAVKVQHPKVRAQSSKDILLMEVLVLAVKQLFP
Q9W133_DROME/137-253    QDLHC.NPEEIFDSFEREPLGTASLAQVHKARLKT...GE.....LVAVKVQHPYVKGNSRVDMKTMELAVNVLARIFP
MCP2_YEAST/166-289      EDLGT.SIEDMFLEFNKTPIGVASLAQVHVAKLKNSDGKGS....SVAVKCQHPSLKEFIPLDVMLTRTVFELLDVFFP
MCP2L_SCHPO/168-286     VDTGK.GLDETFFDEFDPIALGVASLAQVHKARLKDS...DV.....WVAVKVQHPSVSLNSPLDLSMTRWVFKAIKTTFF
Q9VTG5_DROME/162-278    KDFGQ.LPEEIYQEFDYQPVAAASLAQVFKARLPS...GE.....QVAVKVQYNDLQKRFISDLGTIIFLQDIVEFFFK
Y2090_ARATH/147-268     KEVGE.MPDQVFAEFDVPPIASASLAQVHVARTH.D...GK.....KVAVKVQHAAMTDTAAADTAAGVLVNTLHRIFP
YF9E_SCHPO/167-287      EQYGR.PVEEVFASIEKRAAASASIAQVHRAVLPS...GE.....KVAVKIQKPDVAKQMSWDLVYKYMMYVYDKWIF
Y647_MYCTU/150-271      EELGD.EPARLFASFEEEPFASASIAQVHYATLR.S...GE.....EVVVKIQRPGIRRRVAADLQILKRFQOTVELAKL
Q9ZCP5_RICPR/34-153     HSTNTCNVSLPFLHFDNPNPIAAASISQVHKAQLIT...GG.....YVALKILRPDIRKKYNRDIKLLYFFAKIISKFSK
H2VFS0_ZYMMO/111-228    .ALGC.PIEKSFRFFNEIPIGSASIAQVYQAETLD...GV.....TVAVKVLRPGIKLAFRKATETYEWAATKIESLND
Q89WD1_BRADU/112-231    .SLER.PLKDVFASLGP.PVAAASIAQVHRGEVVRD...GIR...KAVAVKVLRPNVASRFRRDLSDFFVVAHKAETYS
Y445_PBCV1/90-208       EKES..PPDFTFYEWYKEPIASASIATVYKGRKTD...NS.....DVILKVRVPEVKQRIMEDLPLFTIIVLDIAKFFGV
UBIB_SHIDS/115-232      ...GL.PVEAWFDDFEIKPLASASIAQVHTARLKS...GK.....EVVIKVRIPDILPVIKADLKLIYRLARWVPRLLP
```

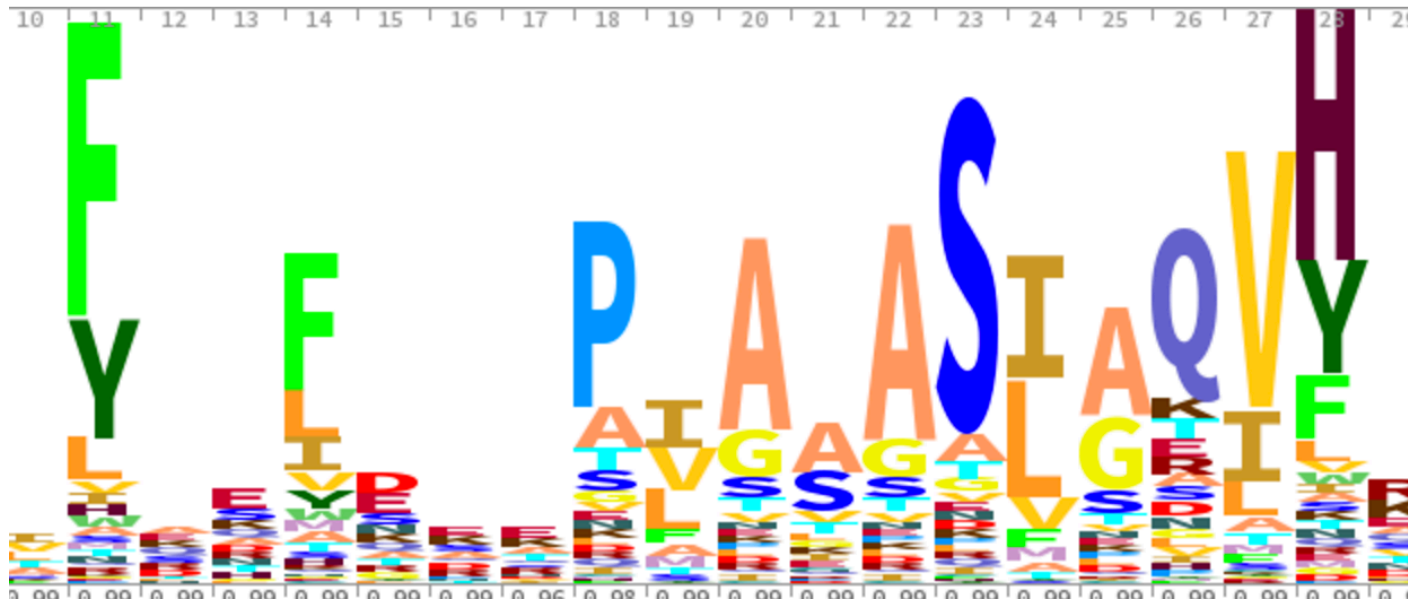

Becomes a Profile

- 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

ALCLIMLAHSGGGAISPVSNTTRTPNLPTINDSKKQIENSTTTTPPPTTQDQSYSCSVCNKAFASYQALGGHKASHRKNATATASDDG
NHSTSTSTTTAAASTASNVSALNPRGRLHECSICHKSFTPTGQALGGHKRRHYEGIIGGGSSKSSVTSSDGGASSHAPRDFDLNLPATP
EFQLELTVDCVKKSQFVG DQEVE SPMPFKKPRT.PT.FGERF

- Results

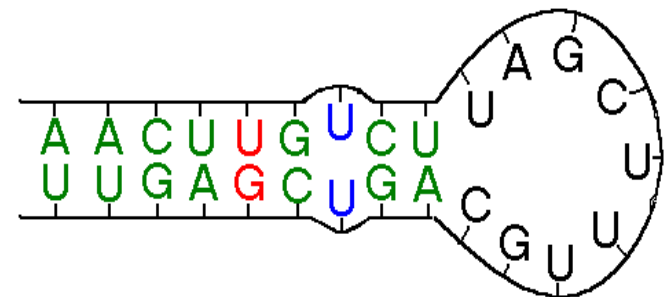
Family	Description
zf-C2H2_6	C2H2-type zinc finger
#HMM	heCdeCsksfpSlqaLggHkksHrk
#MATCH	+ C++C+k F S+qaLggHk+sHrk
#PP	78*****8
#SEQ	YSCSVCNKAFASYQALGGHKASHRK

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

Infernal + Rfam



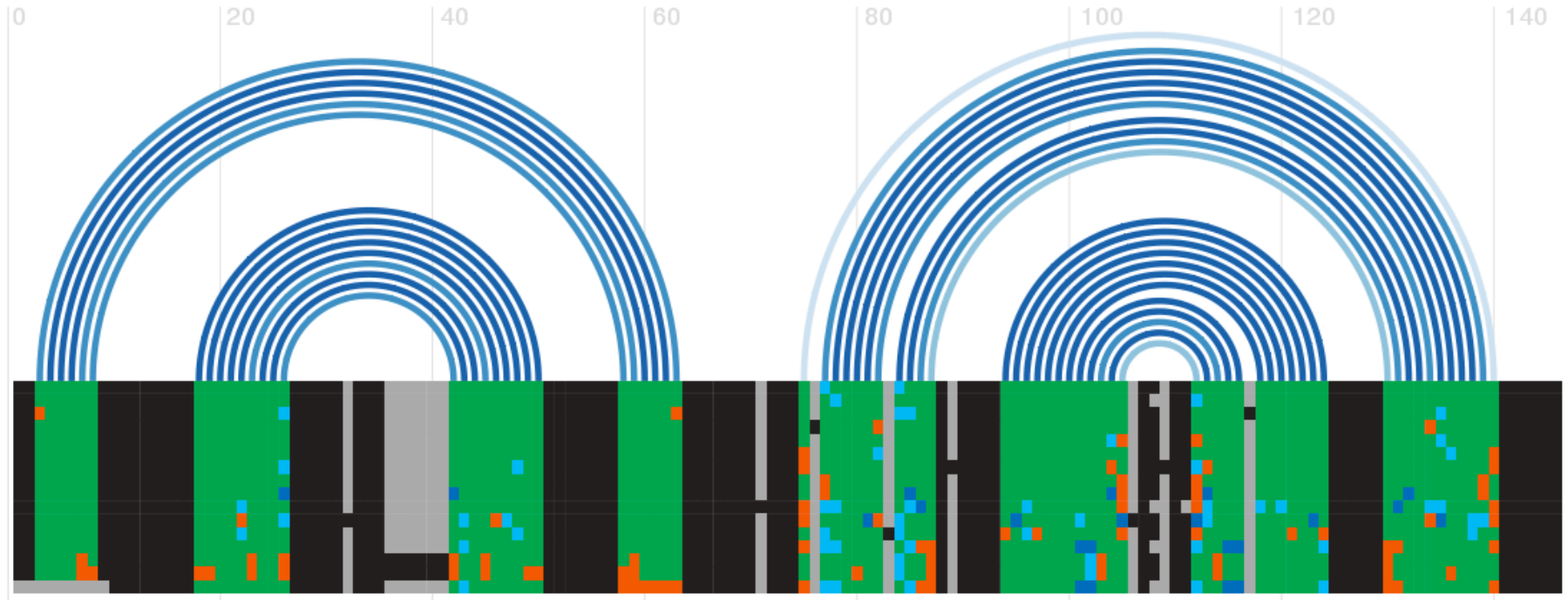
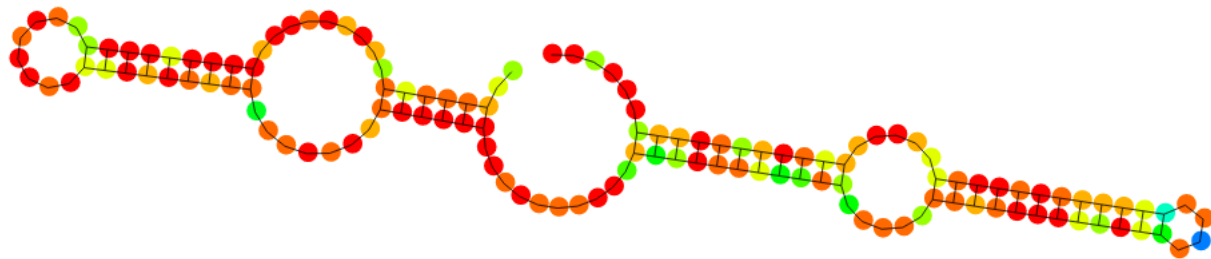
- Infernal ("INFERence of RNA ALignment")
- Tool for searching for RNA structure and sequence similarities
- Uses information about sequence AND structure
- Also uses HMM (generates covariance model instead of a profile)
- Start with:
 - Multiple sequence alignment
 - Special annotation of bases that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



Example snoRNA

AAAGCAGGUUGCAAUACAGUGCUUCAUUUU.GUG.....GAAGUACUGCCAUAUCCUGCGUGAAAGAA.AAGC.CGUGUU.AAUCA.UUUUUGAUUUUGCCUU.UA.
AAAGCAGGUAGCAAUACAGUGCUUCAUUUU.GUG.....GGAGUACUGCCAUAUCCUGCGUGAAAGAA.AAGC.CAUGUU.GGUUG.UUUCUGAUUUUUGCUU.U-
-----AGCAAUACAGUGCUUCAUUUU.GUG.....GGAGUACUGCCAUAUCCUGCGUGAAAGAA.AAGC.CAUGUU.GGUUG.UUUCUGAUUUUUGCUU.U-
AGAGCAGGUUGCAAUACAGUGCUUCAUUUU.GUG.....GAAGUACUGCCAUAUCCUGCGUGAAAGAA.AAGC.UAUGUU.GAUCU.UUUUUGAUUUUGCCUU.C-
AAAGCAGGUUGCAAUACAGUGCUUCAUUUU.GUG.....GAAGUACUGCCAUAUCCUGCGUGAAAGAA.AAGC.UGUGUU.GAUCG.UUAUUGAUUUUGCCCA.UA.
AAAGCAGGUUGCAAUACAGUGCUUCGUUUU.GUG.....GAAGUACUGACAUAUCCUGCGUGAAAGAA.AAUC.AGUGUU.GAUCU.UUUUUGAUUUUGCCUC.UC.
AAAGCAGGUUGCAAUACAGUGCUUUCUUUU.GUG.....GAAGUAUUGACAUAUCCUGCGUGAAAGAA.AAUC.UGUGUU.GAUCG.UUUUUGAUUUUGCCAU.UUa
UACGCAGGUUGAAUACAGUGCUUUGUUUU.GGG.....GAAGUACUGCUGUAUCCUGCGUGAAAGAC.AAGC.UGUGUU.AGUCA.UUUUUGAUUUUGCCUU.UA.
AAAGCAAGCUGCAAUACAGUGCUUCAUUUU.GUGaaauacUAAAUACUGCCAUAUCCUGCGUGAAAGAA.AAGC.UGUGUU.AAUGA.UUUUUGAUUUUGUCUU.UG.
AAAGCAGGCUGCAAUACAGUACUUCAGUUU.GUG.....GAAGUACUGCCAUAUCCUGCGUGAGAGAAaAAGC.CAUGUU.GGCCG.GCUCUGGUUUUGCCUC.U-
UGAGCAGGUUGCAGUUCAGUCCUUUGUUUUcGUG.....GGAGUGCUGGCAUAACCCUGCGUGAAAACA.AAUA.UGUGCC.AAUCA.UUUUUUAAUUUACCUCaUU.
UAAGCAGGUUGCAAUACAGUGCUUCAUUUU.GUG.....GAAGUACUGACAUAUCCUGCGUGAAAGAA.AAUCuUGUGUG.GAUCU.UUUUUGAUUUUGCCUU.UG.
AAAGCAGGUUGCAAUACAGUACUUCAUUCU.GUG.....GAAGUAUUGCCAUAUCCUGCGUGAAAGAA.AAGC.CGUGUUuAAUCA.UUUCGGGUUUUGCCUG.UA.
AAAGCAGGUUGCAAUACAGUGCUUCAUUUU.GUG.....GAAGUACUGACAUAUCCUGCGUGAAAGAA.AAUG.UGUGUC.GAUCU.UUUUUGAUUUUGCCUU.UA.
AAAGCAAGCUGGAAUUGCAGUGCUUCAUUUU.GUGaaauacUAAAUACCAUCAUAUAGCUGCGUGAAAGAA.AAGC.UGUUUU.AAUGA.UUUUUGAUUUUGUCUU.UG.
AAAGCAGGUUGCAAUACAGUGCUUUAUUUU.GUG.....AAAGUACUGUCAUAUCCUGCGUGAAAGAA.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 *

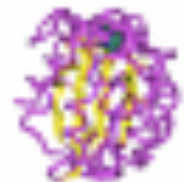
InterProScan

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:

- Inflorescence
- 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

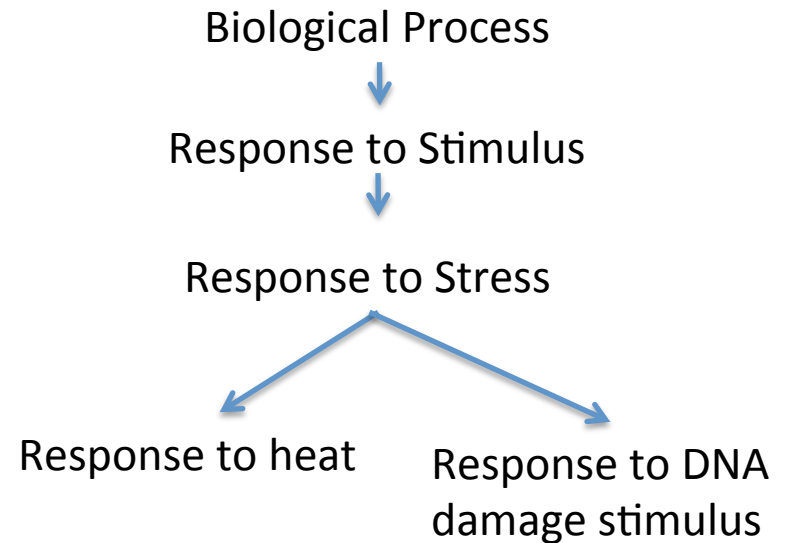
- All these would be coded in a computer readable structure:
 - Inflorescence
 - Flower
 - Gynoecium (Pistil)
 - Androecium
 - Perianth



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

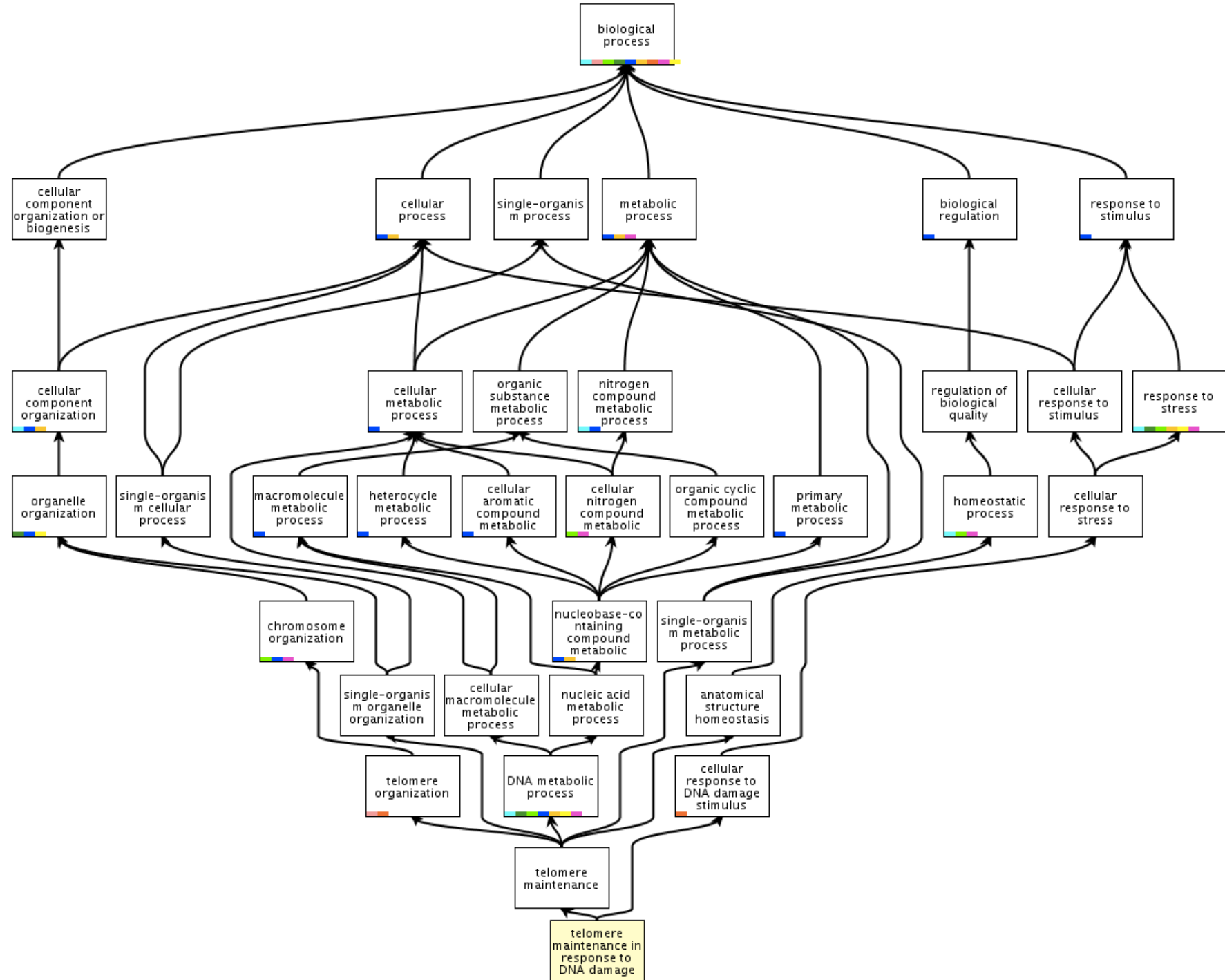
GO:2001020 regulation of response to DNA damage stimulus

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

