Functional Annotation

Functional vs Structural Annotation

- Structure:
  - Location of exons/introns/UTRs
  - Location of regulatory elements
  - Location of repeats
  - Identification of genomic elements by position
- Function:
  - What does it do?
  - Biological implications
  - Genes - what proteins and pathways?
  - Regulatory elements – what do they regulate?

Outline

- Hidden Markov Models
- HMMER search of Pfam
- Infernal search of Rfam
- Gene Ontology

Pfam and HMMER

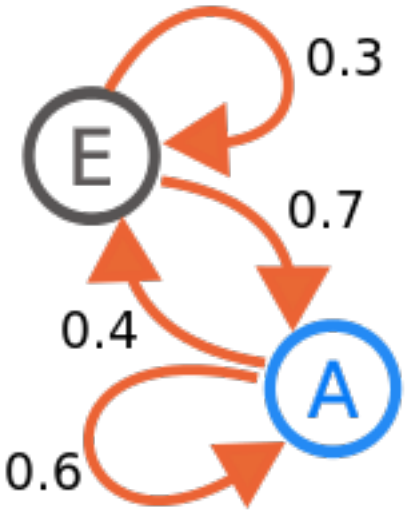
Pfam + HMMER

- 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…

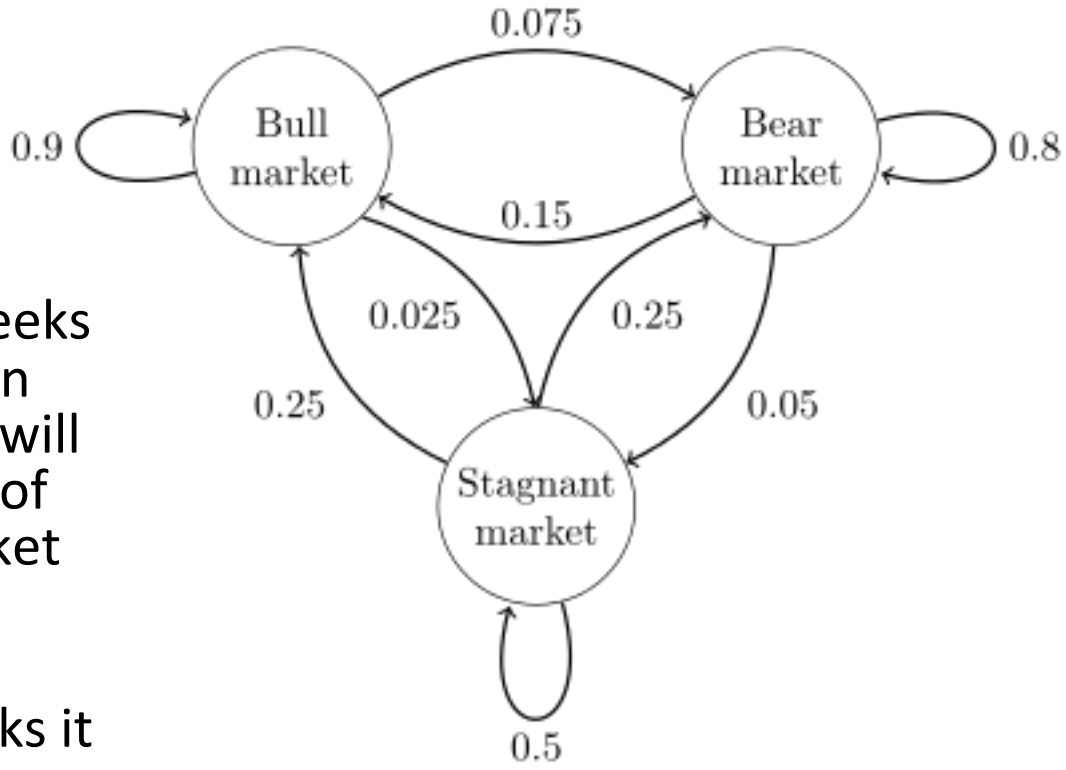
Markov Chain

- 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

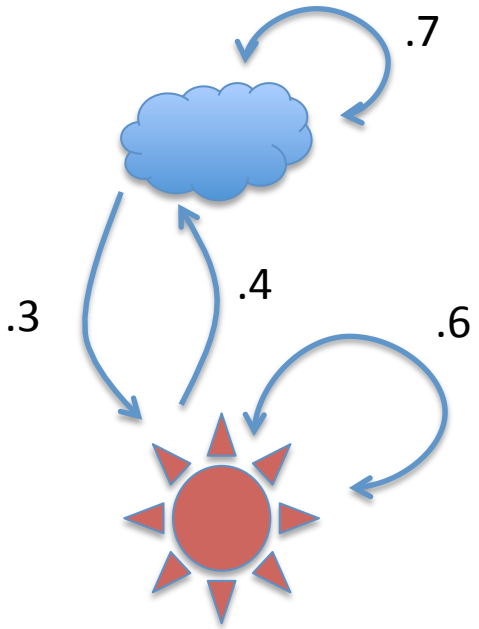
Hidden Markov model

- 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's city weather

- If the day is Rainy
- 70% chance the next day will be rainy
- 30% change the next day will be sunny
- If the day is Sunny
- 40% chance the next day will be rainy
- 60% cance the next day will be sunny
-



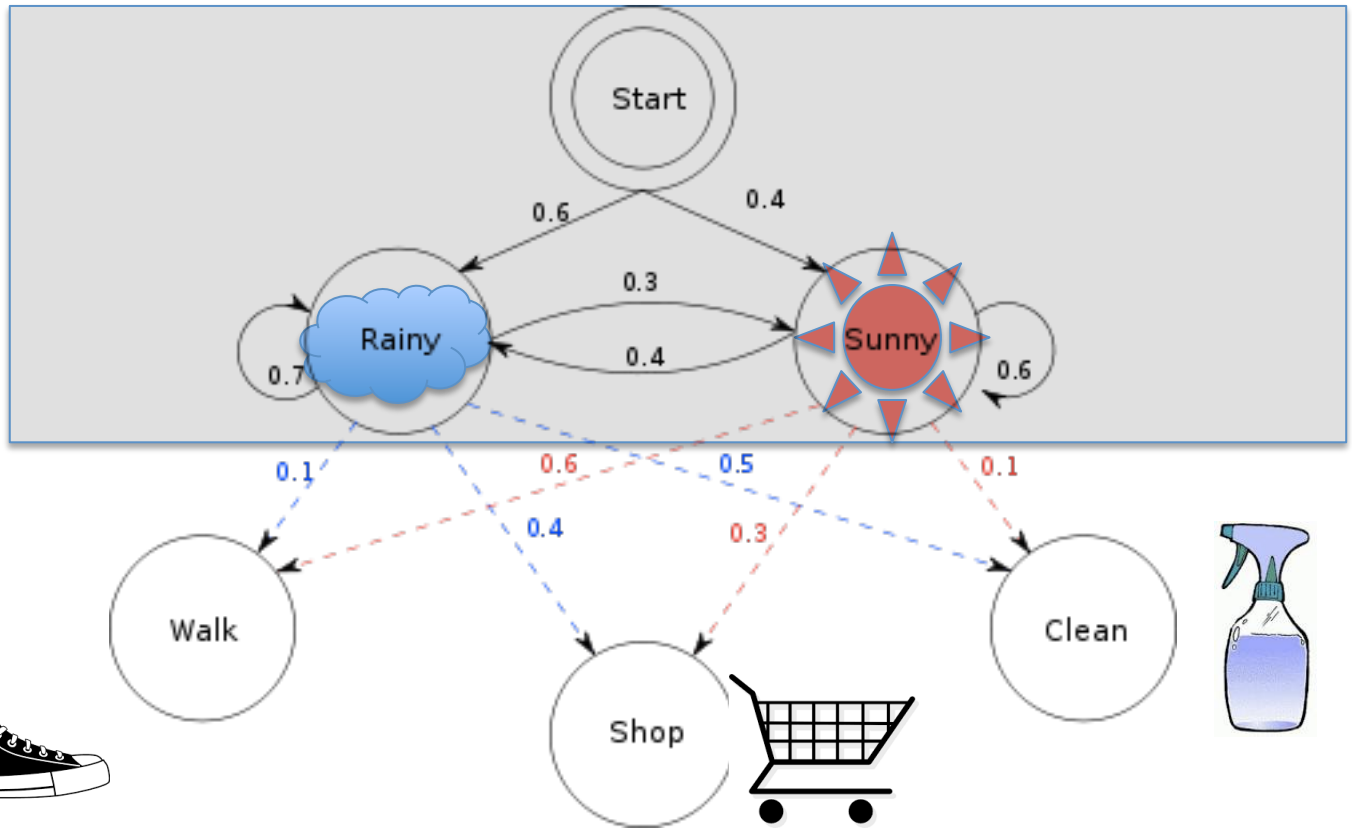
HIDDEN MARKOV MODEL

- Bob decides what to do based on the weather. He only has three activities: walking, shopping and cleaning.

- If its rainy
- 10% change he will walk
- 40% chance he will shop
- 50% chance he will clean
- If its sunny
- 60% change he will walk
- 30% chance he will shop
- 10% chance he will clean

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.



Weather state hidden from Susan

Known by Susan

HIDDEN MARKOV MODEL – The 3 problems

Using a string of observations about Bob's behavior, what can Susan deduce?

- {walk, clean, clean, shop, walk, walk}

1. Given the model and observations, what is the likelihood of this sequence of observations occurring?
2. Given the model and observations, what is the likeliest hidden state sequence to have produced this sequence of observations?
3. Given the observations and the number of hidden states (but not knowing the model), what is the most likely model? (ie Training a model to best fit the observed data.)

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

How should the parameters be set?
How do we weight them?
How to score?
How confident that an answer is correct?

- HMMs provide a conceptual toolkit for buidling complex models.

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

????????

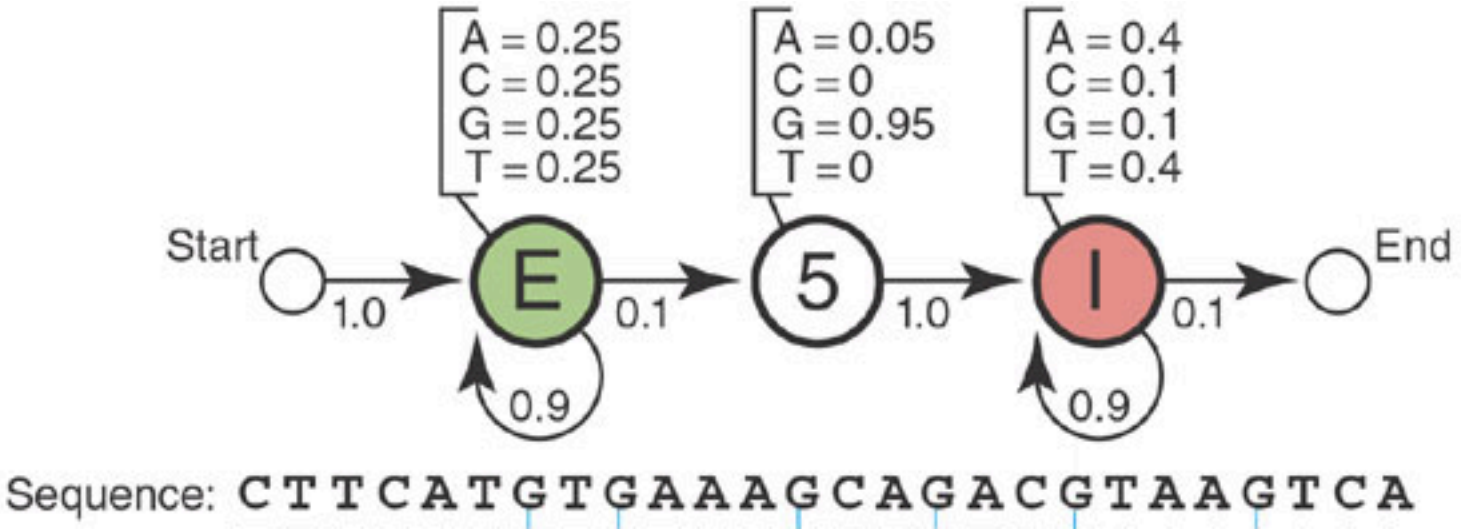
Where is the splice site?

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

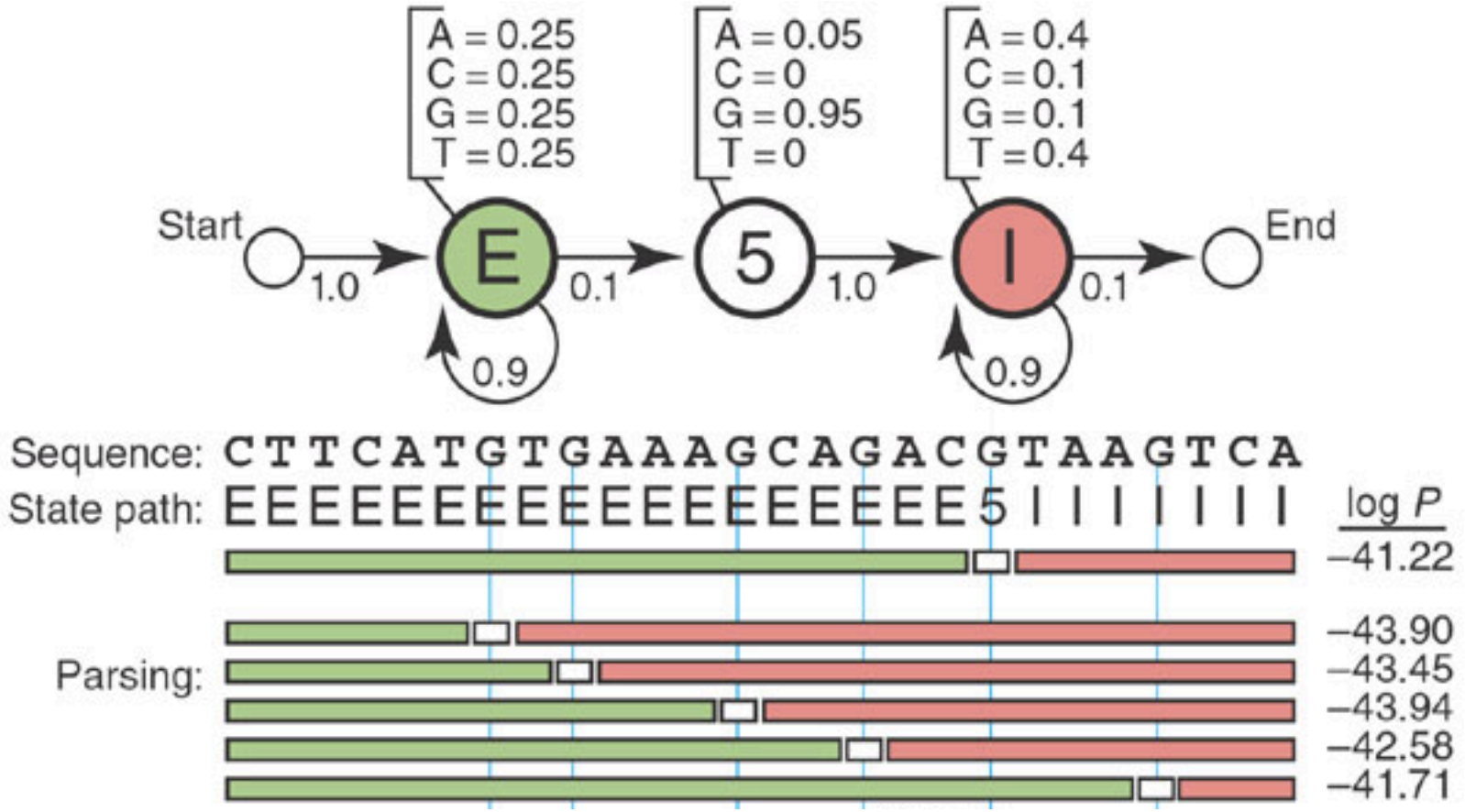
HMM – 5’ splice example



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

Eddy
What is a hidden markov model?
2004

HMM – 5’ splice example



Eddy
What is a hidden markov model?
2004

- 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?

Pfam

- 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

Pfam

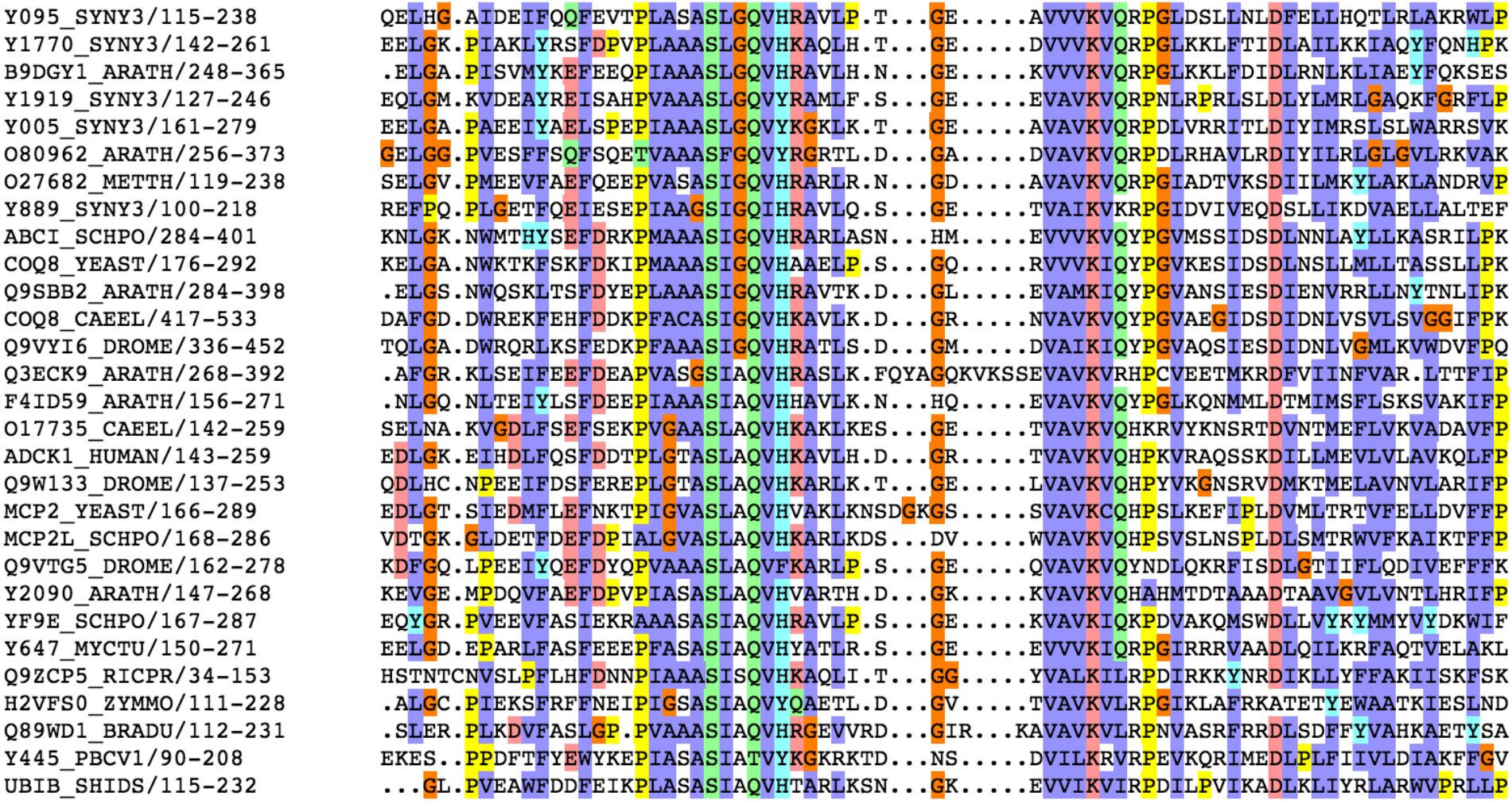
- Currently has 16,306 families (version 30)
- Families are grouped into “clans” - related by similarity of sequence, structure or profile-HMM
- Family information includes gene architecture, structure, sequences from hundreds to thousands of species and interactions.

Pfam

- HMMER is used by Pfam for two purposes:
  - To construct Pfam clusters (build the model)
  - To detect matches from a given sequence to a cluster (compare a sequence to see how well it fits the model)

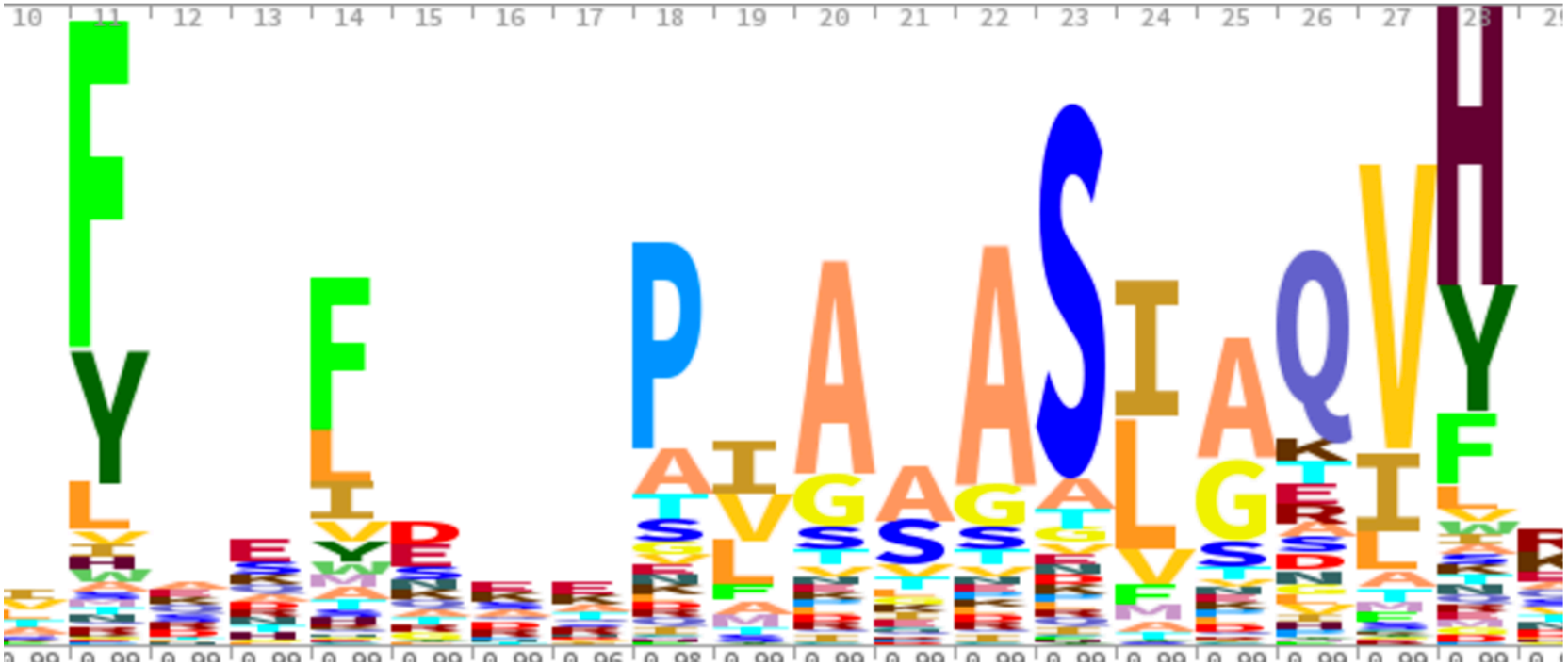
- The matching amino acid probabilities from the HMM can be visualized as a logo

Multiple Sequence Alignment



Becomes a Profile

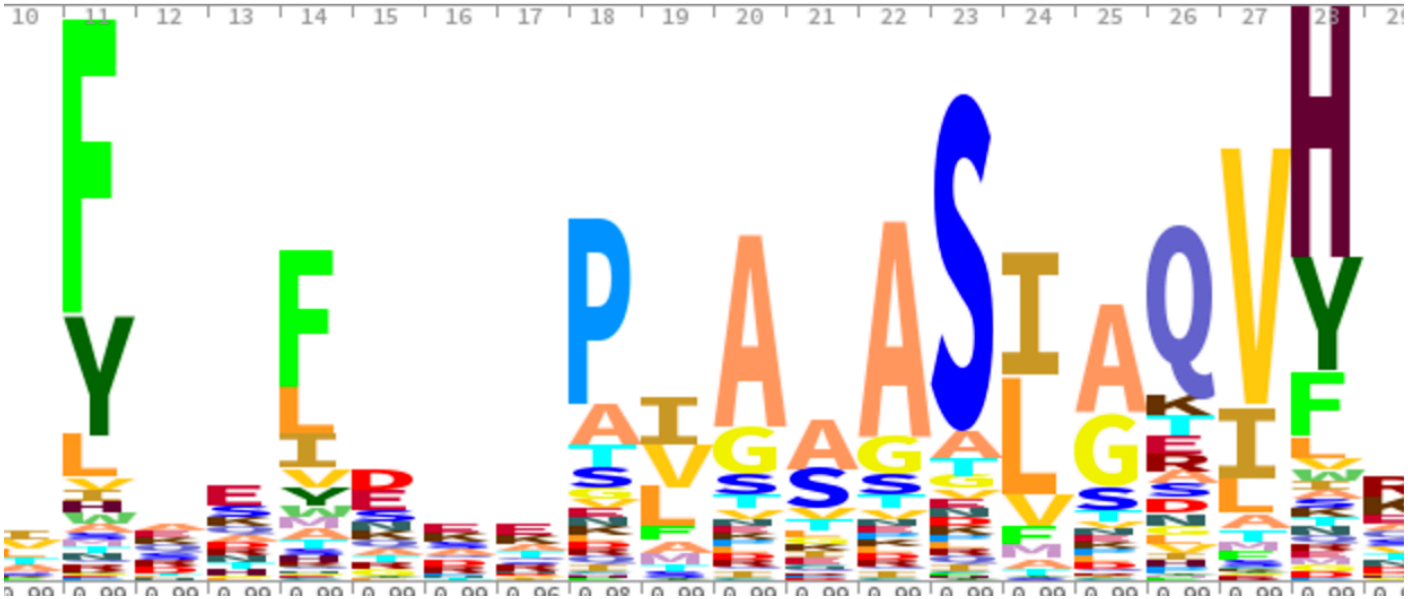
- HMM represented as a logo



Height = letter probability

Pfam

- Take all 16,306 profiles
- Use hmmsearch to compare your sequences to these families
- E-values –just like BLAST



Example:

- Search green ash protein

```
ALCLIMLAHSGGGAAISPSVNTTRTPNLPTINDSKKQIENSTTTPPPTTQDQSYSCSVCNKAFASYQALGGHKASHRKNATATASDDG
NHSTSTSTTTAAASTASNVSALNPRGRLHECSICHKSFPTGQALGGHKRRHYEGIIGGGSSKSSVTSSDGGASSHAPRDFDLNLPATP
EFQLELTVDCVKKSQFVGDQEVESPMPFKKPRLPLFGERF
```

- Results

Family	Description
zf-C2H2_6	C2H2-type zinc finger

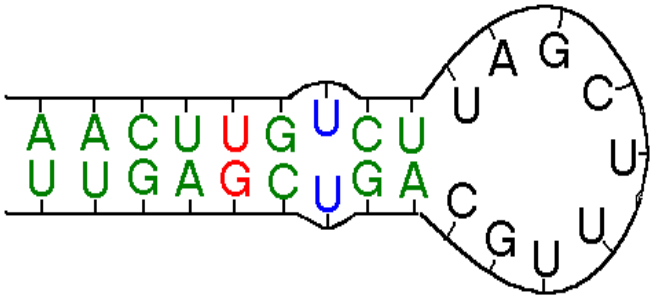
```
#HMM     heCdeCsksFpSlqaLggHkksHrk
#MATCH   + C++C+k F S+qaLggHk+sHrk
#PP      78**********************8
#SEQ     YSCSVCNKAFASYQALGGHKASHRK
```

	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

- Start with:
  - Multiple sequence alignment
  - Special annotation of bases that are paired to create the secondary structure



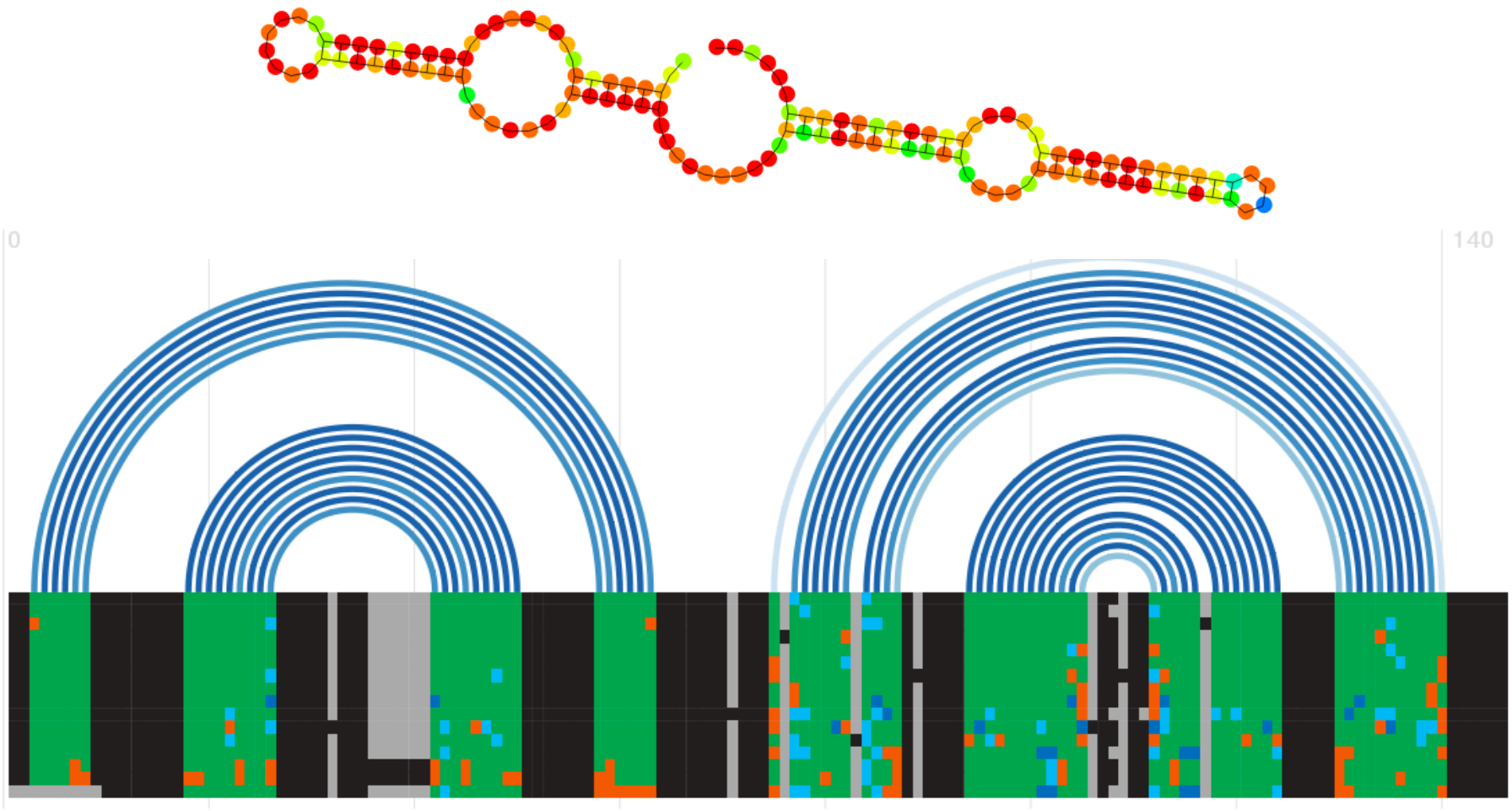
Rfam

Version 12.1
2,474 families

- Contains RNA families with conserved secondary structure
- 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).
- Search with Infernal

Example snoRNA

Example snoRNA



Hidden markov model uses information about sequence conservation and structure conservation

Ontology

Ontology

- Roots in philosophy – how we conceptualize and specify knowledge (Aristotle)
- Very useful for organizing information with 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
    - Syrah (Shiraz)

From NCBI SRA for Arabidopsis

I want sequences that relate to flower structures. I have to manually look through a list and select:

- Inflorescense
- Inflorescence
- Immature inflorescence
- Flower
- Flowers
- Pistils pollinated for 8 Hours
- 3xHA_infloresence_biological_replication1
- 3xHA_infloresence_biological_replication2
- 3xHA-VvCEB1-OX_inflorescence_biological_replication3

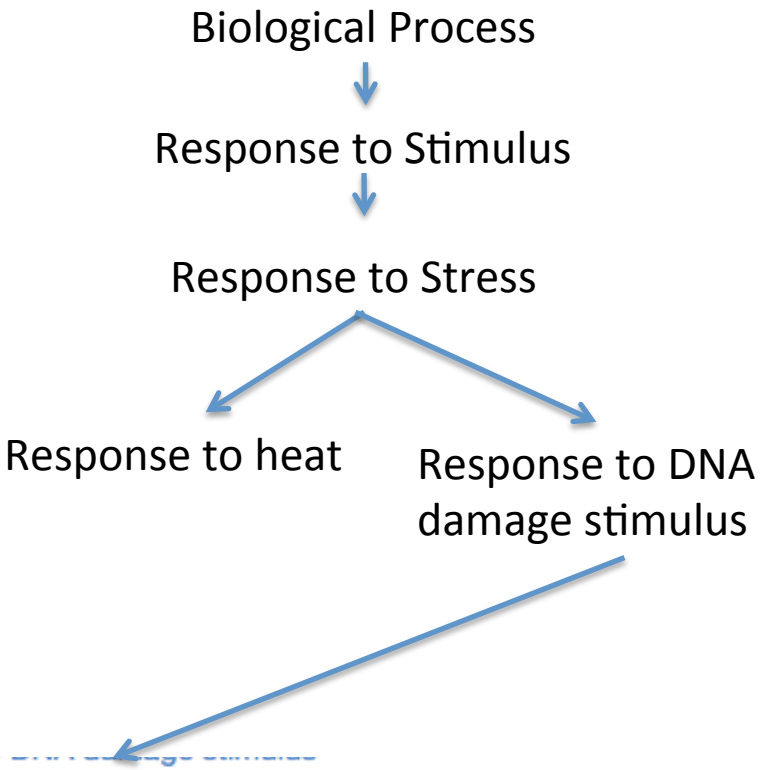
Plant Structure Ontology

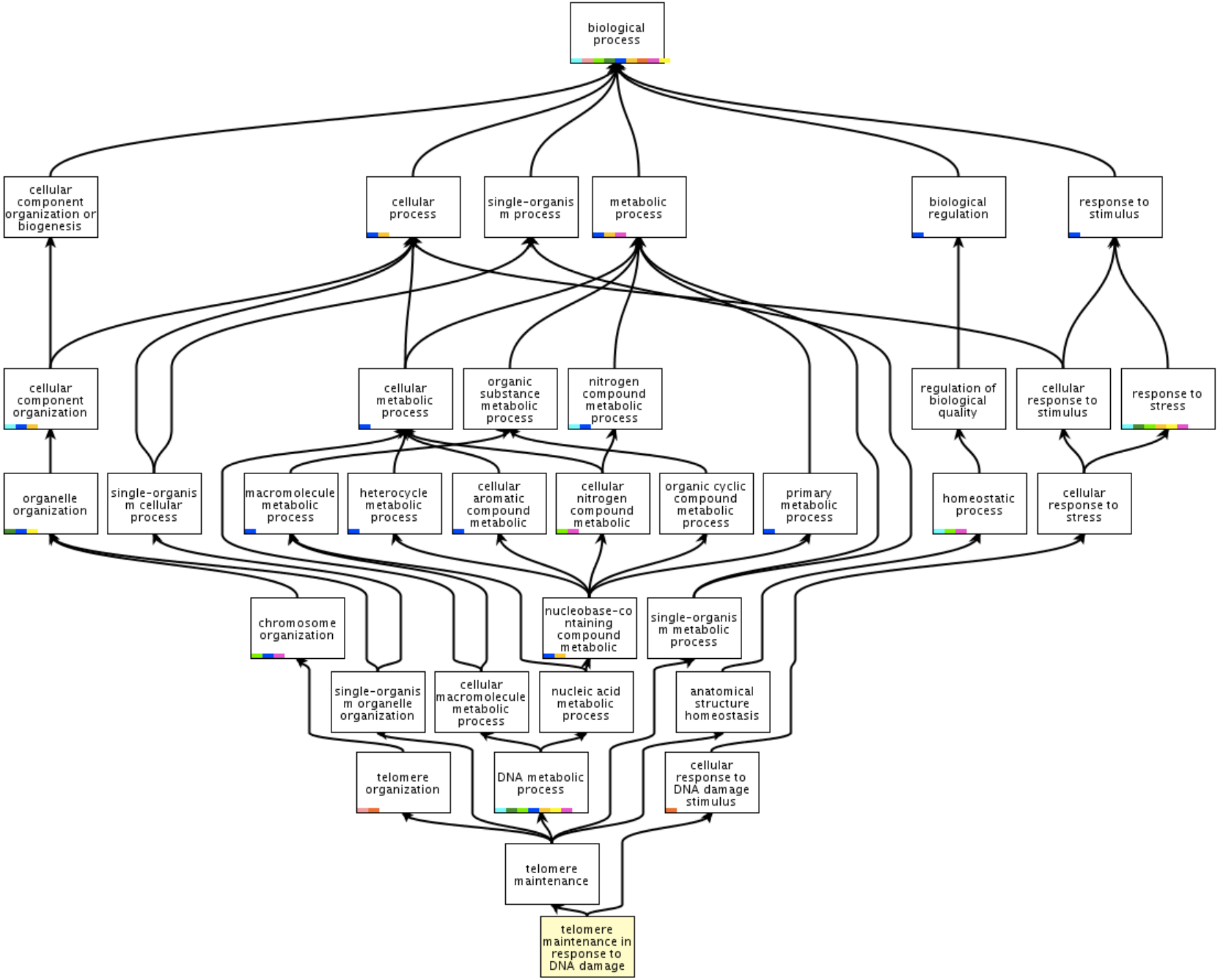
- All these would be coded in a computer readable structure:

  - 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




biological process
cellular component organization or biogenesis
cellular process
single-organism process
metabolic process
biological regulation
response to stimulus
cellular component organization
cellular metabolic process
organic substance metabolic process
nitrogen compound metabolic process
regulation of biological quality
cellular response to stimulus
response to stress
organelle organization
single-organism cellular process
macromolecule metabolic process
heterocycle metabolic process
cellular aromatic compound metabolic
cellular nitrogen compound metabolic
organic cyclic compound metabolic process
primary metabolic process
homeostatic process
cellular response to stress
chromosome organization
nucleobase-containing compound metabolic
single-organism metabolic process
single-organism organelle organization
cellular macromolecule metabolic process
nucleic acid metabolic process
anatomical structure homeostasis
telomere organization
DNA metabolic process
cellular response to DNA damage stimulus
telomere maintenance
telomere maintenance in response to DNA damage

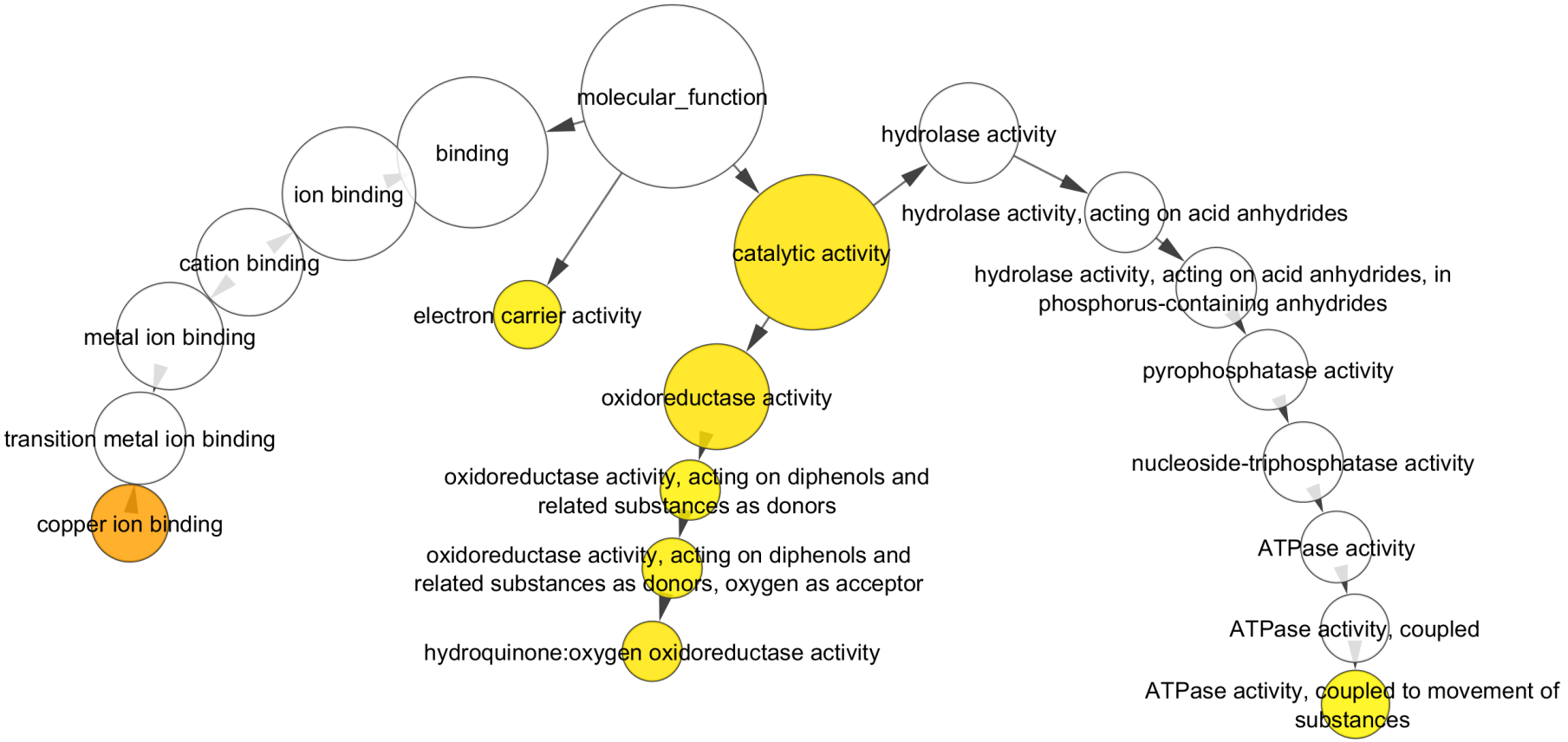
Uses of Gene Ontology

- Example (Without an ontology)
  - You search for genes that are induced by stress
  - The computer searches through a database of genes
  - Some are annotated “response to drought”
  - The computer does not return them because it does not see the word “stress” anywhere
- Example (With an ontology)
  - You search for genes that are induced by stress
  - The computer searches through a database of genes
  - Some are annotated “response to drought”
  - The computer looks that up in our vocabulary and figures out that drought is a type of abiotic stress, which is a type of stress
  - The computer now returns those genes to you
- Knowledge has become more accessible!!!

Uses of Gene Ontology

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

Example


molecular_function
binding
ion binding
cation binding
metal ion binding
transition metal ion binding
copper ion binding
electron carrier activity
catalytic activity
oxidoreductase activity
oxidoreductase activity, acting on diphenols and related substances as donors
oxidoreductase activity, acting on diphenols and related substances as donors, oxygen as acceptor
hydroquinone:oxygen oxidoreductase activity
hydrolase activity
hydrolase activity, acting on acid anhydrides
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
pyrophosphatase activity
nucleoside-triphosphatase activity
ATPase activity
ATPase activity, coupled
ATPase activity, coupled to movement of substances