Genome Annotation and Sequence Alignment

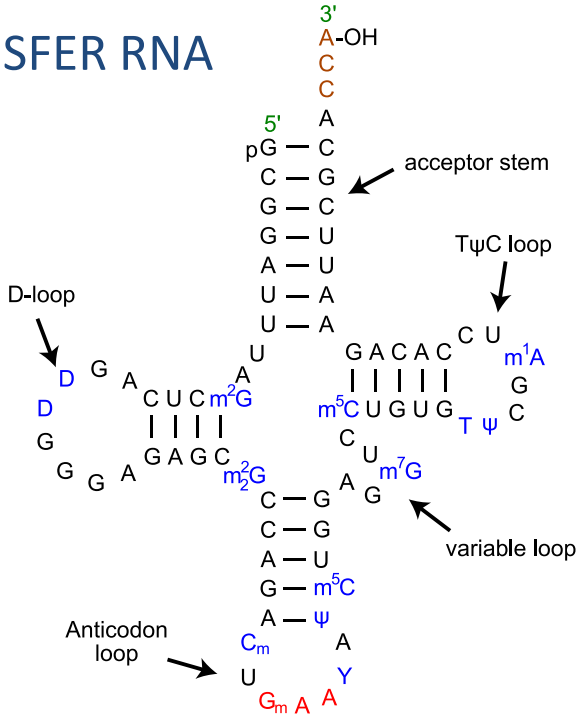
DAVID L. TABB, PH.D.

Overview

- Hidden Markov Models find protein-coding genes. RNA-Seq confirms them.
- Sequence alignment algorithms are essential tools; substitution matrices are their “score cards.”
- Matching known sequence motifs suggests similar domains and thus shared function.

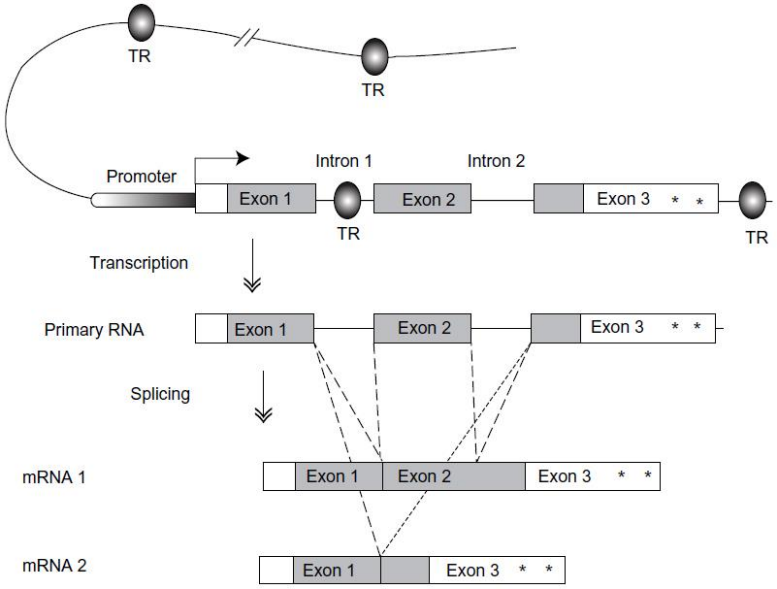
What does a gene look like?

TRANSFER RNA



tRNA-Phe from *S. cerevisiae*, Wikimedia Commons Yikrazuul

PROTEIN-CODING GENE

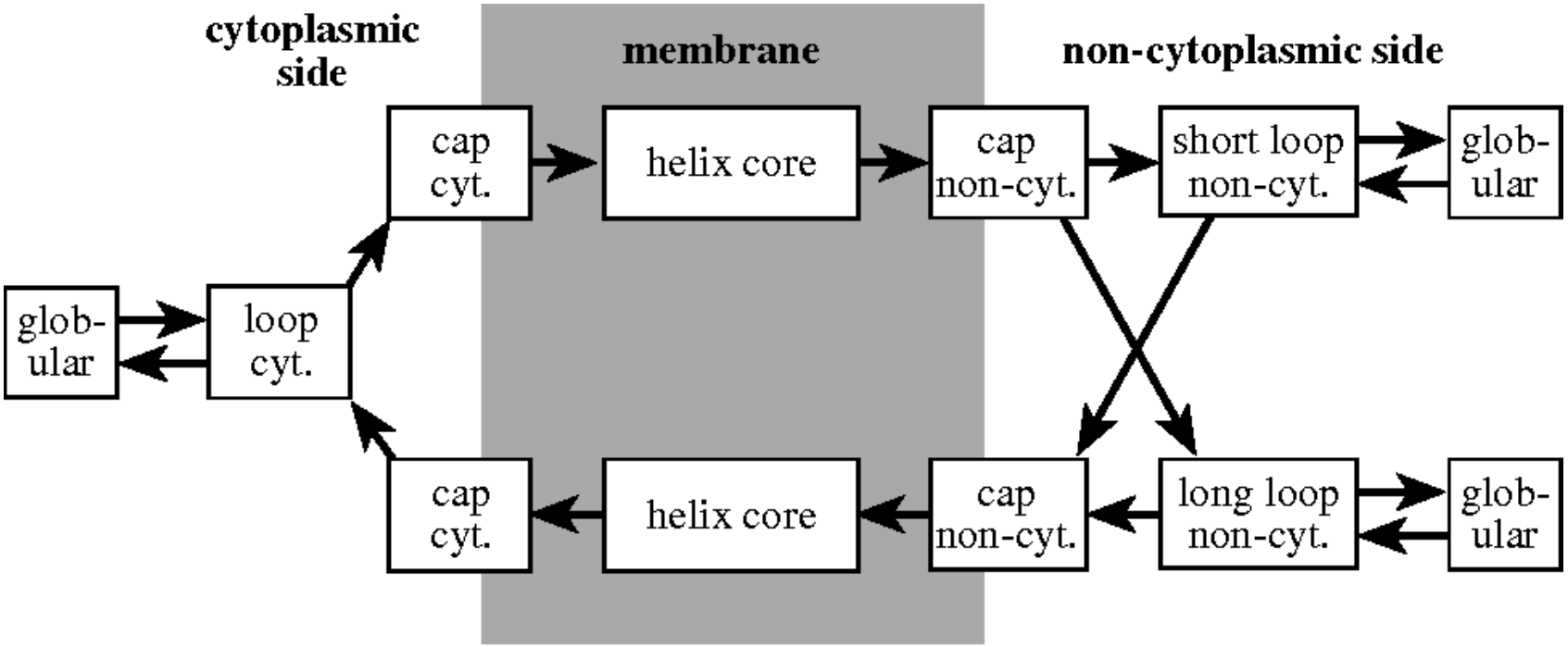


L. Sastre. *Adv. in Genomics and Genetics* (2014) 4: 15-27.

Today's topic: finding protein-coding genes.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Hidden Markov Models represent different bio features as “states”



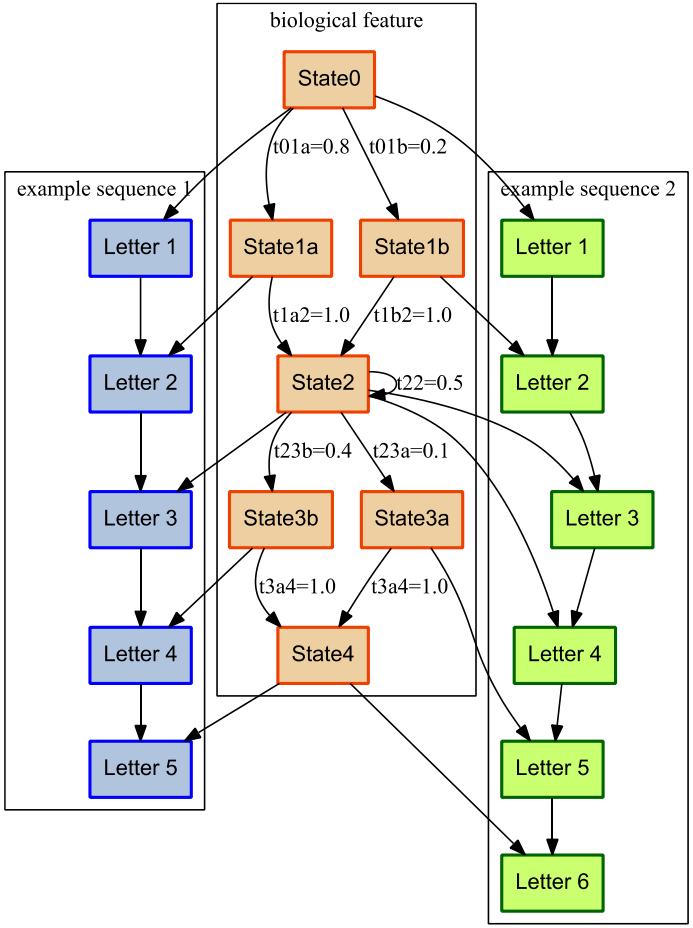
High-level depiction of Transmembrane HMM
https://services.healthtech.dtu.dk/service.php?TMHMM-2.0

Sonnhammer ELL et al. *ISMB-98 Proceedings* (1998) 6: 175-182.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Each state of an HMM "emits" sequence. We move between states by "transitions."

- Transition probabilities: bio features comprise many connected states, with several possible paths.
- Emission probabilities: each state yields characteristic sequences

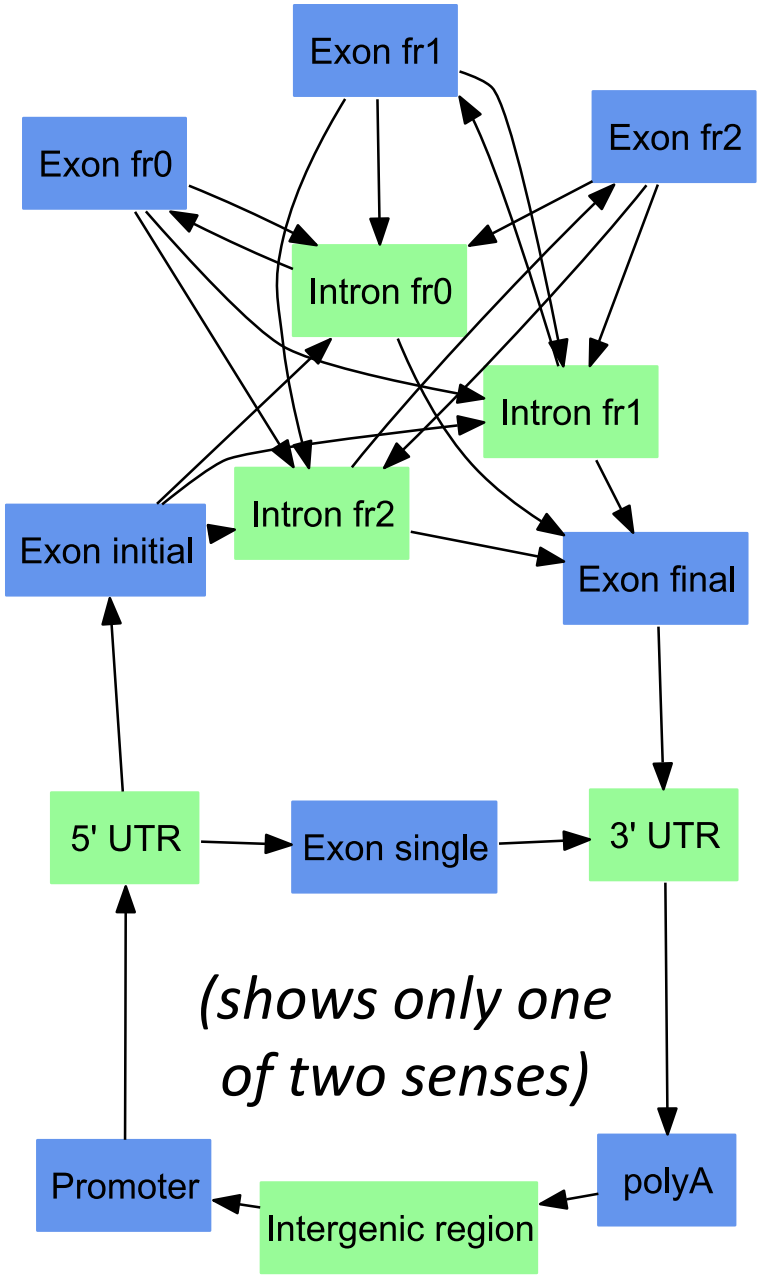


Key algorithms train HMMs and fit sequences to them.

- Baum-Welch: Train an HMM from a dataset of sequences. Finds the most likely set of state transition and emission probabilities.
- Forward-backward: Score a new sequence against HMM by computing the best probability that this sequence would result from model.
- Viterbi: Associate parts of the test sequence to components of this biological feature. Where do we step from state to state in sequence?

GENSCAN HMM

- Finds positive and negative sense genes simultaneously
- Incorporates transcriptional, splicing, and translational signals in its model
- Detects multiple genes within long genomic contigs



Burge and Karlin. *J. Mol. Biol*. (1997) 78-94.

"computational"

UNIVERSITEIT STELLENBOSCH UNIVERSITY

How well do *in silico* methods support gene finding?

"from scratch"

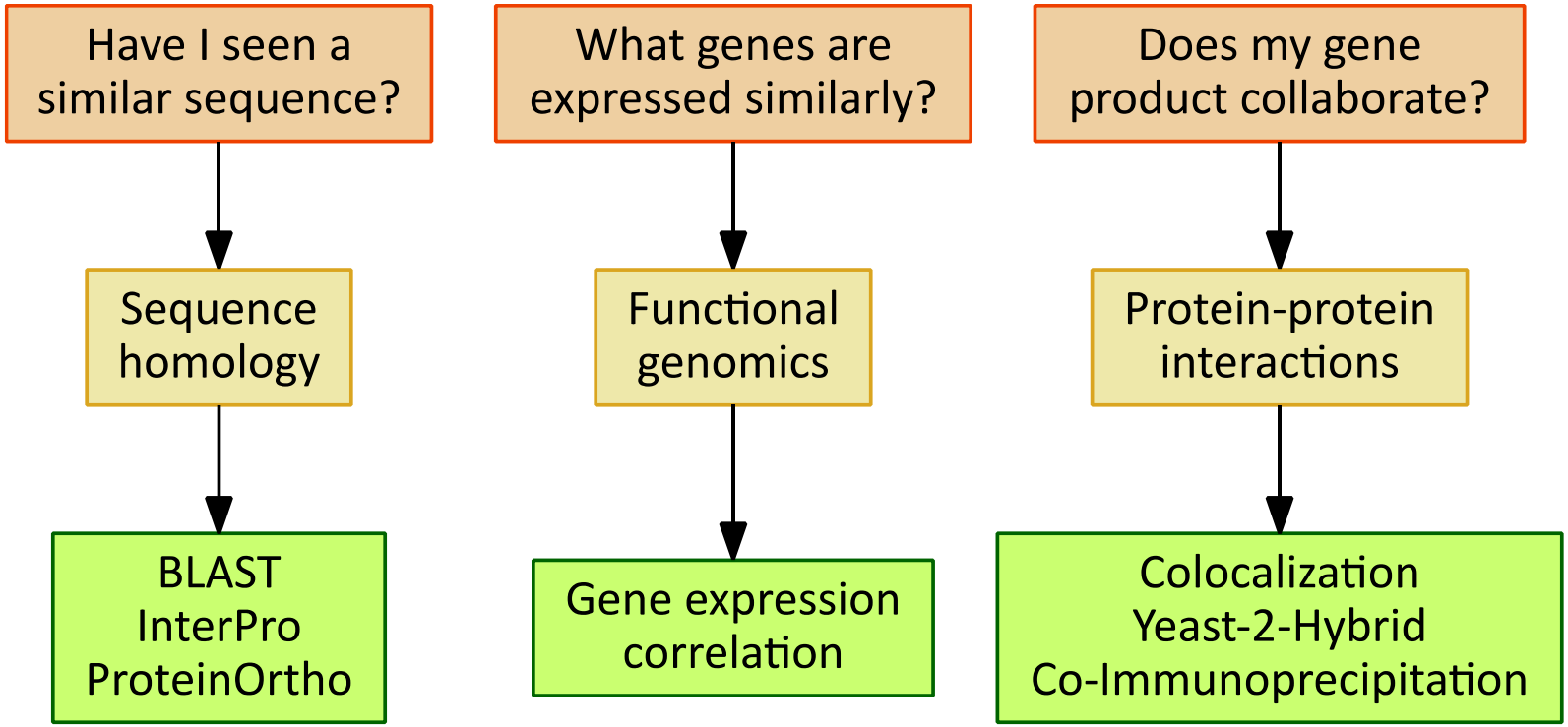
- *Ab initio* HMM gene finders are prone to finding exons where they do not exist.
- Similarity-based gene prediction tools depend upon closely-related orthologs.
- Synteny (same gene ordering in closely related species) can assist gene annotation.
- RNA-Seq experiments are now routinely used for gene model confirmation.

R Guigó et al. *Genome Research* (2000) 10:1631-1642.

M. Alexandersson et al. Genome Research (2003) 13: 496-502.

Intermission

What does this gene do?

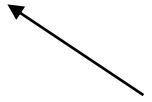


Why align sequences?

- *Recognize orthologs*: Having newly sequenced an organism, find genes that match known genes in other organisms.
- *Recognize paralogs*: Determine whether a sequenced gene is part of a gene family within this genome.
- *Recognize conserved regions*: Find segments of conserved sequence that may have functional or evolutionary significance.

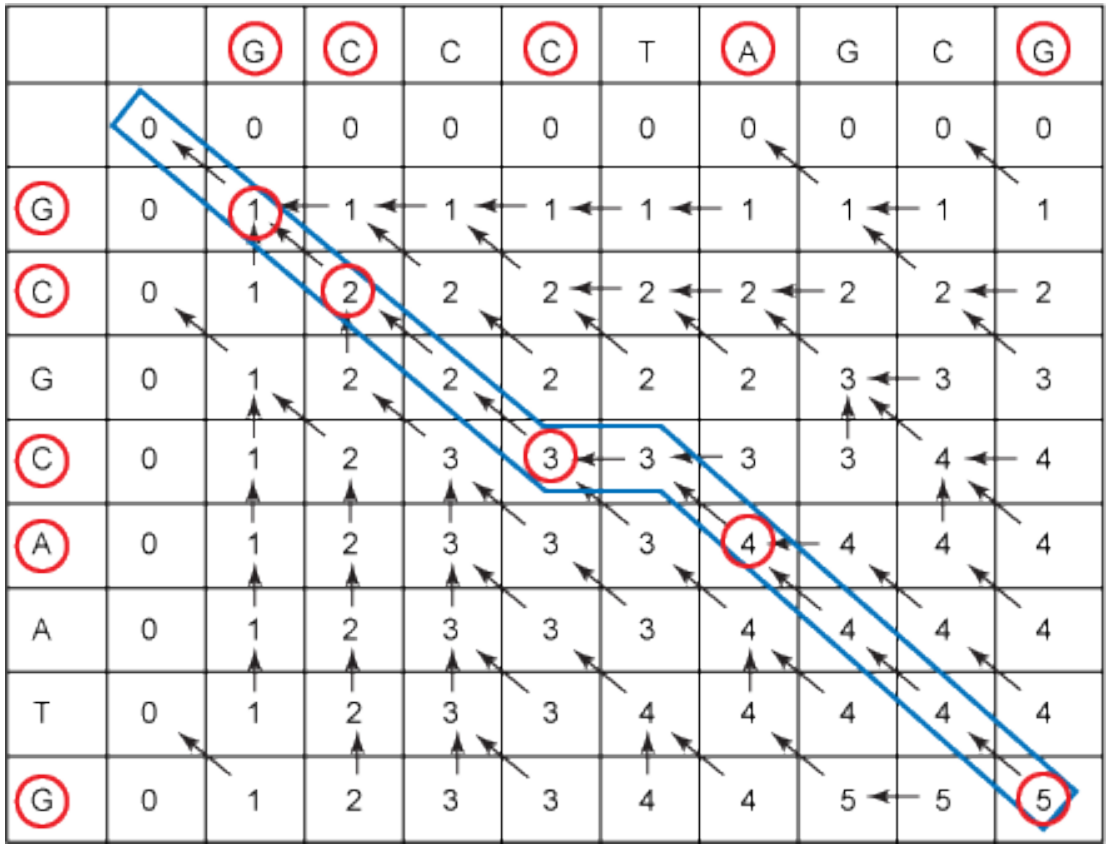
High-level view of alignment

- Global vs. local: Are we aligning whole sequences or finding best internal matches?
- Optimal vs *heuristic*: Is a provably best result required, or is a great one sufficient?
- Gaps may be penalized two different ways:
  - Penalty for opening a gap, and
  - Penalty for extending a gap.



"Affine Gap Penalties"

Smith-Waterman (local) and Needleman-Wunsch (global)



		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1
C	0	1	2	2	2	2	2	2	2	2
G	0	1	2	2	2	2	2	3	3	3
C	0	1	2	3	3	3	3	3	4	4
A	0	1	2	3	3	3	4	4	4	4
A	0	1	2	3	3	3	4	4	4	4
T	0	1	2	3	3	4	4	4	4	4
G	0	1	2	3	3	4	4	5	5	5

- *Dynamic programming* makes best alignment an O(mn) problem.
- Yields provably best result for given scoring.
- May incorporate gap initiation and extension penalties.

https://www.ibm.com/developerworks/library/j-seqalign/

DNA vs protein alignment

- *Identity-based scoring:* perfect matches +1, but what about transitions / transversions?
- *Substitution matrix*: using amino acid sequences requires a “score card” for each possible replacement.
- DNA sequences can be compared for evolutionary neighbors, but proteins are needed to detect distant relationships.

Position-specific scoring matrices

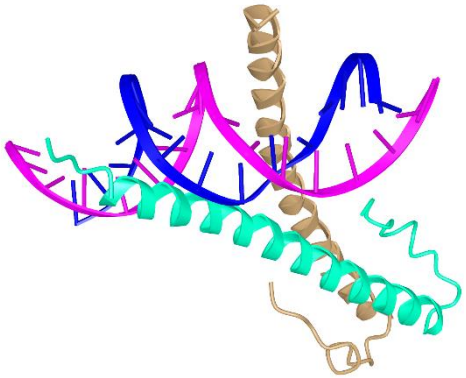
- A sequence motif can be described as a PSSM.
- Each column of a PSSM defines a particular position in a biological sequence.
- Each row of a PSSM gives a score for a particular letter.
- A PSSM is similar to an HMM in having different emission probabilities, but there is only one series of moves through its “states.”

UNIVERSITEIT STELLENBOSCH UNIVERSITY

cd14809: bZIP_AUREO-like

basic leucine zipper

Dimerization of leucine zippers creates a pair of adjacent basic regions that bind DNA and undergo conformational change.

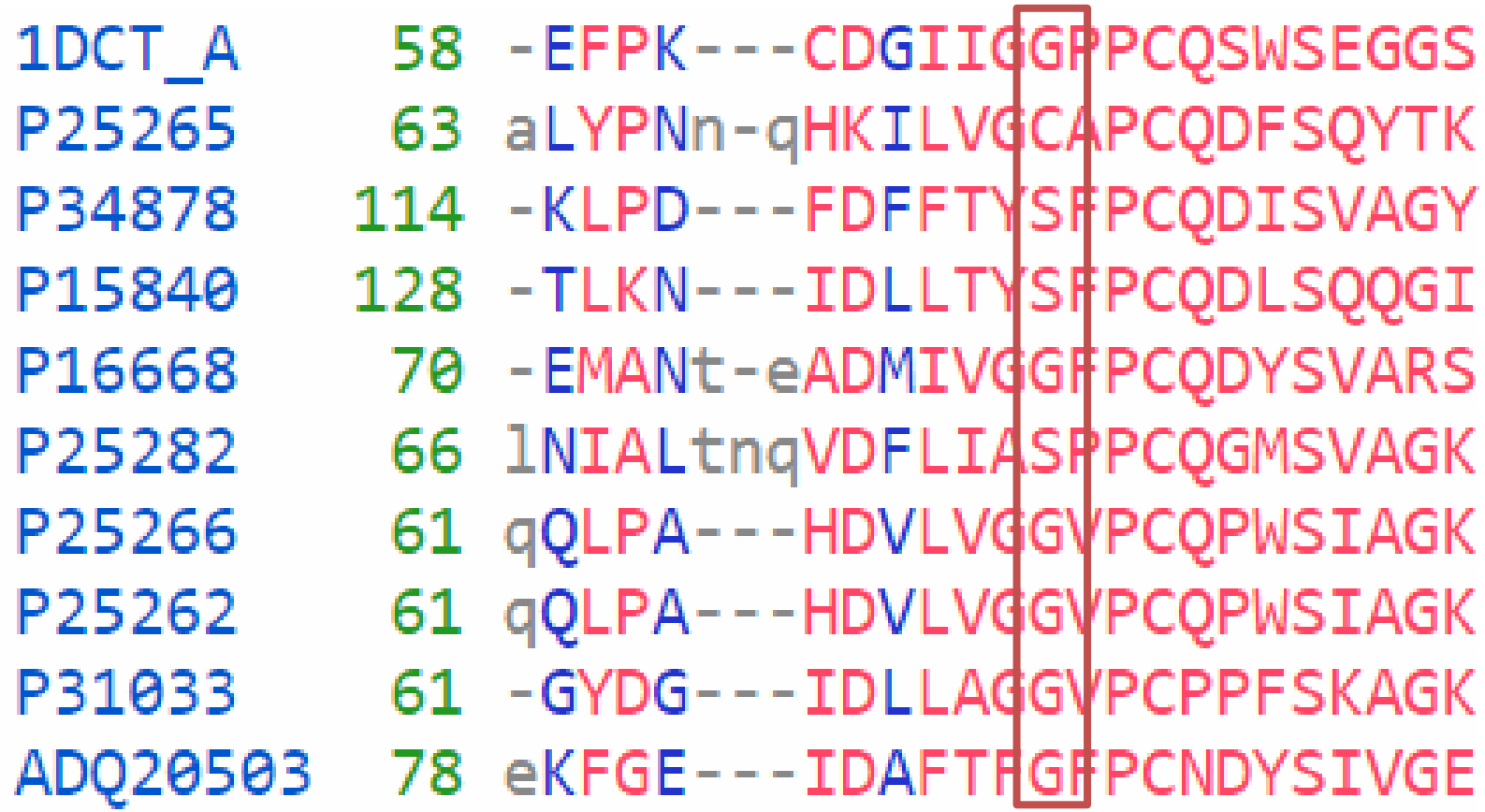


```
Feature 1                  # ### ###### ###
2C9L_Y          2 LEIKRYKNRVAARKSRAKFKQLLQHYREVAAAKSSENDRLRLLLKQMCPSLD 53
2C9L_Z          2 LEIKRYKNRVAARKSRAKFKQLLQHYREVAAAKSSENDRLRLLLKQMCPSLD 53
gi 300256532  791 SWQRRERNRLAARKCRAKKMEFIAHMQAQLDAMAARNEELRVQLVALQRMYS 842
gi 300267855  443 ARERREKNRIAARKCRAKKVAMVRGMQEELKELVAQNQEMKMQVVTWHRMFS 494
gi 24943134   180 QERKRYRNRLASRRCRAKFRNQLEHFRTVAAAKTEENNRLRVLIRQMCPTLD 231
gi 332016862  470 PKTRKEKNKLASRACRLKKKAQHEANKIKLHGLETEHRRLIQGISQAKHTLA 521
gi 154693802  229 REKYLEKNRRAARKCRLKKKAEMAADQAKYDHYVSELRATKKQLEDSRKELL 280
gi 242345219   62 KEKRKERNKLLARKSRMKLKADLENLKAKLMYLMKENESLRSQLYRVSTPPV 113
gi 323450651  755 TALERERNREHARNTRARKKEAIEKLKHDVEAWEVETRRTEERRAVKERKSE 806
gi 323452895   70 AELTRRRNRENARSTRMRRKMYIKHLQQVAETLKERHDELNSKMAAPPPDGY 121
```

https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=cd14809

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Which AA can replace another and survive natural selection?



S. Henikoff and JG Henikoff. *PNAS* (1992) 89: 10915-10919

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Two probabilities make an odds ratio

- Observed probability of occurrence:

q_{ij} = the fraction of table sum found in this cell.

- Expected probability of occurrence:

e_{ij} = the product of the background probabilities of either residue.

- BLOSUM rounds values of $\log_2(q_{ij}/e_{ij})$.
- 0 is expected rate, positive is more than usual.

S. Henikoff and JG Henikoff. *PNAS* (1992) 89: 10915-10919

Substitution Matrix: BLOSUM62 is our "Score card."

UNIVERSITEIT STELLENBOSCH UNIVERSITY

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

S. Henikoff and JG Henikoff. *PNAS* (1992) 89: 10915-10919

```
R E G A D A V M L S G E T A H G K Y P L
-2+2+6+0+6+0+1+5+4+4+0+5+5+0-3+6+5+7+7+2
Y D G T D C L M L S N E T T I G K Y P I
```

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Intermission

BLAST: the killer app

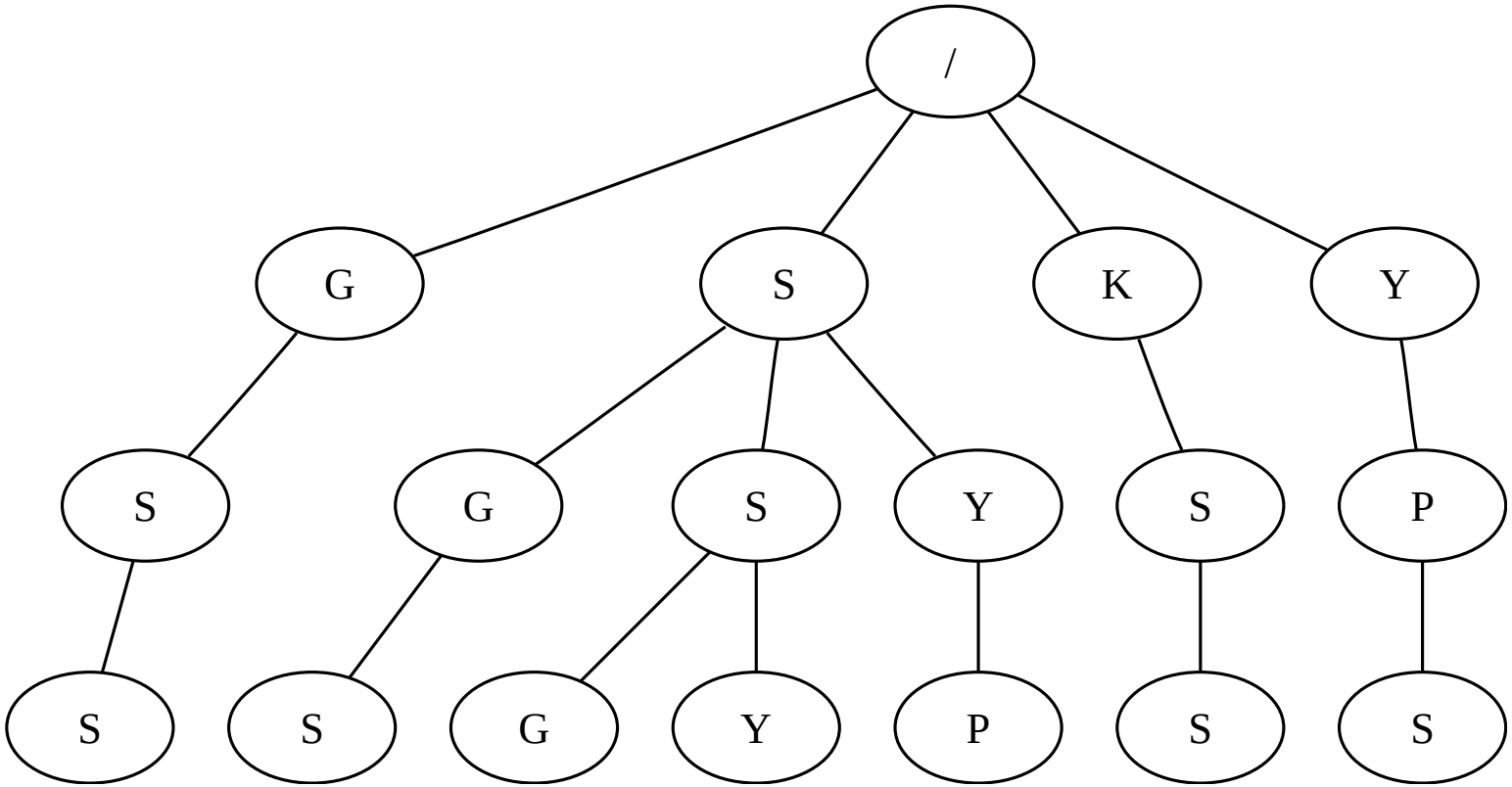
1. Find exact "*seed*" matches between query and DB sequences via a finite state machine.
2. Extend each seed to a *maximal segment pair* (MSP), allowing for approximate matches.
3. Sort all MSPs by alignment score, and display those that are unlikely to be random hits.

Expectation value: The number of alignments with scores at least this good that were expected to occur by chance. Lower is better.

Altschul et al. *J. Mol. Biol*. (1990) 215: 403-410
Karlin and Altschul. *PNAS* (1990) 87: 2264-2268

UNIVERSITEIT STELLENBOSCH UNIVERSITY

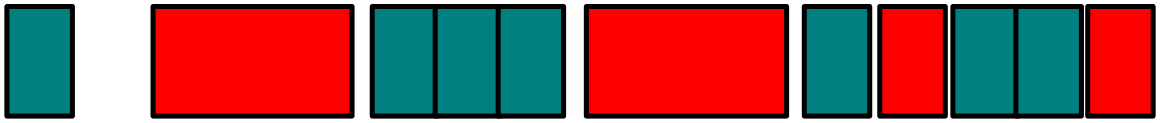
Finite State Machine for KSSGSSYPS



TLLAYNKSMAVWTAYRIIASGSPRDLHADETELYWS

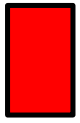
Extending from a seed to a Maximal Segment Pair

PFERPEAEAMCTSFKENPT



RLVRPEVDVMCTAFHDNEE

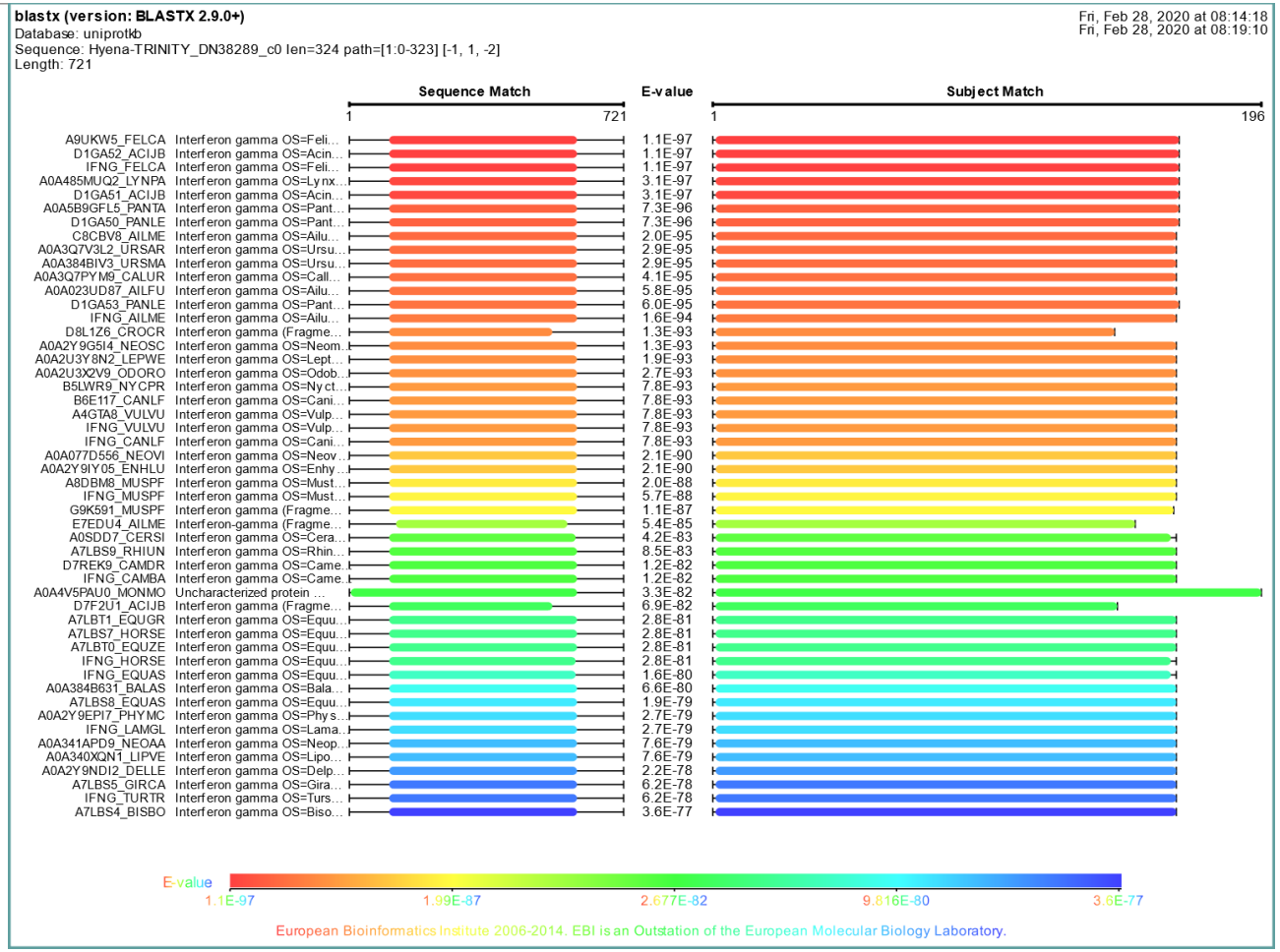
 Seed matches

 Identical residues

 Similar residues

BLAST can be run on each sequence or in batch.

- blastx matches mRNA to protein.
- E-values show strength of match.
- Taxonomy shown in accession suffix:
  - FELCA: housecat
  - ACIJB: cheetah
  - LYNPA: lynx
  - PANTA: tiger



https://www.ebi.ac.uk/Tools/sss/ncbiblast

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Why use Multiple Sequence Alignment?

- We use BLAST to match a query sequence in a large database. MSA seeks relationships among a user-supplied set of sequences.
- When the sequences that are most similar to each other are closely related in evolutionary time, the MSA can recapitulate a ***phylogeny***.
- MSA helps us to learn which residues in sequences are most ***conserved*** over evolutionary time.

JD Thompson et al. *Nucl. Acids. Res*. (1994) 22: 4673-4680

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Clustal Omega MSA Orthologs

```
sp|Q9LIK0|PKP1_ARATH      DSL---TNLEEIILASDGAMVARGDLGAQIPLEQVPAAQQRIVQVCRALNKPVIVASQLL 405
sp|P00549|KPYK1_YEAST     QGV---NNFDEILKVTDGVMVARGDLGIEIPAPEVLAVQKKLIAKSNLAGKPVICATQML 301
sp|O62619|KPYK_DROME      QGM---HNLDEIIEAGDGIMVARGDLGIEIPAEKVFLAQKAMIARCNKAGKPVICATQML 334
sp|P30613|KPYR_HUMAN      EGV---KRFDEILEVSDGIMVARGDLGIEIPAEKVFLAQKMMIGRCNLAGKPVVCATQML 374
sp|P53657|KPYR_MOUSE      EGV---KKFDEILEVSDGIMMARGDLGIEIPAEKVFLAQKMMIGRCNLAGKPVVCATQML 374
sp|P21599|KPYK2_ECOLI     EAVCSQDAMDDIILASDVVMVARGDLGVEIGDPELVGIQKALIRRARQLNRAVITATQMM 287
sp|P9WKE5|KPYK_MYCTU      EAI---DNLEAIVLAFDAVMVARGDLGVELPLEEVPLVQKRAIQMARENAKPVIVATQML 279
                          :.:     :: *: . *  *:****** ::   ::   *:  :  ..   : *: *:*::
```

Pyruvate kinase [EC 2.7.1.40] sequences from plants, yeast, insects, mammals, and bacteria show high homology. They are *orthologs*, genes that have a common ancestor but that diverged through speciation.

N Schormann et al. *Protein Sci*. (2019) 28: 1771-1784.

Clustal Omega MSA Paralogs

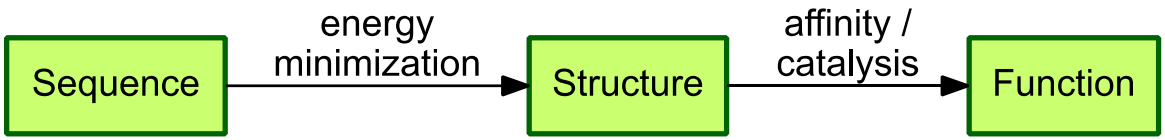
```
sp|P09087|ABDB_DROME      ------------------SPGGL-----RGYPSENYSSSGASGGLSVGAVGPCTPNPGLH  379
sp|P17482|HXB9_HUMAN      VL--------------------------SNQRPG--YGDNKICEGSEDKERPDQTNPSA  177
sp|P31269|HXA9_HUMAN      ----------------------------KQPSEGAFSENNAENESGGDKPPIDPNNPAA  198
sp|P31274|HXC9_HUMAN      ------------------PGELRDRAPQTLPSPE--ADALAGSKHKEEKADLDPSNPVA  184
sp|P28356|HXD9_HUMAN      AAATGGTGPGAGIGAATGTGGSSEPSACSDHPIPGCSLKEEEKQHSQPQQQQLDPNNPAA  277
                                                                                .

sp|P09087|ABDB_DROME      EWTGQVSVRKKRKPYSKFQTLELEKEFLFNAYVSKQKRWELARNLQLTERQVKIWFQNRR  439
sp|P17482|HXB9_HUMAN      NWLHARSSRKKRCPYTKYQTLELEKEFLFNMYLTRDRRHEVARLLNLSERQVKIWFQNRR  237
sp|P31269|HXA9_HUMAN      NWLHARSTRKKRCPYTKHQTLELEKEFLFNMYLTRDRRYEVARLLNLTERQVKIWFQNRR  258
sp|P31274|HXC9_HUMAN      NWIHARSTRKKRCPYTKYQTLELEKEFLFNMYLTRDRRYEVARVLNLTERQVKIWFQNRR  244
sp|P28356|HXD9_HUMAN      NWIHARSTRKKRCPYTKYQTLELEKEFLFNMYLTRDRRYEVARILNLTERQVKIWFQNRR  337
                          :*    * **** **:*.*********** *:::::* *:** *:*:************
```

HOX (homeobox) protein Abd-B is a single gene in drosophila, but it has duplicated twice over in humans to create *paralogs*.

TRJ Lappin et al. *Ulster Med J*. (2006) 75: 23-31.

“Central Dogma” of Structural Biology



- A sequence motif is a conserved element of a protein sequence alignment that usually correlates with a particular function.
- A structural domain is an element of overall structure within a protein that is self-stabilizing and often folds independently of the rest of the protein chain.
- A protein family is a group of evolutionarily, and perhaps functionally, related proteins.

Wright and Dyson. *J. Molec. Biol*. (1999) 293: 321-331.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Integrating multiple methods for motifs

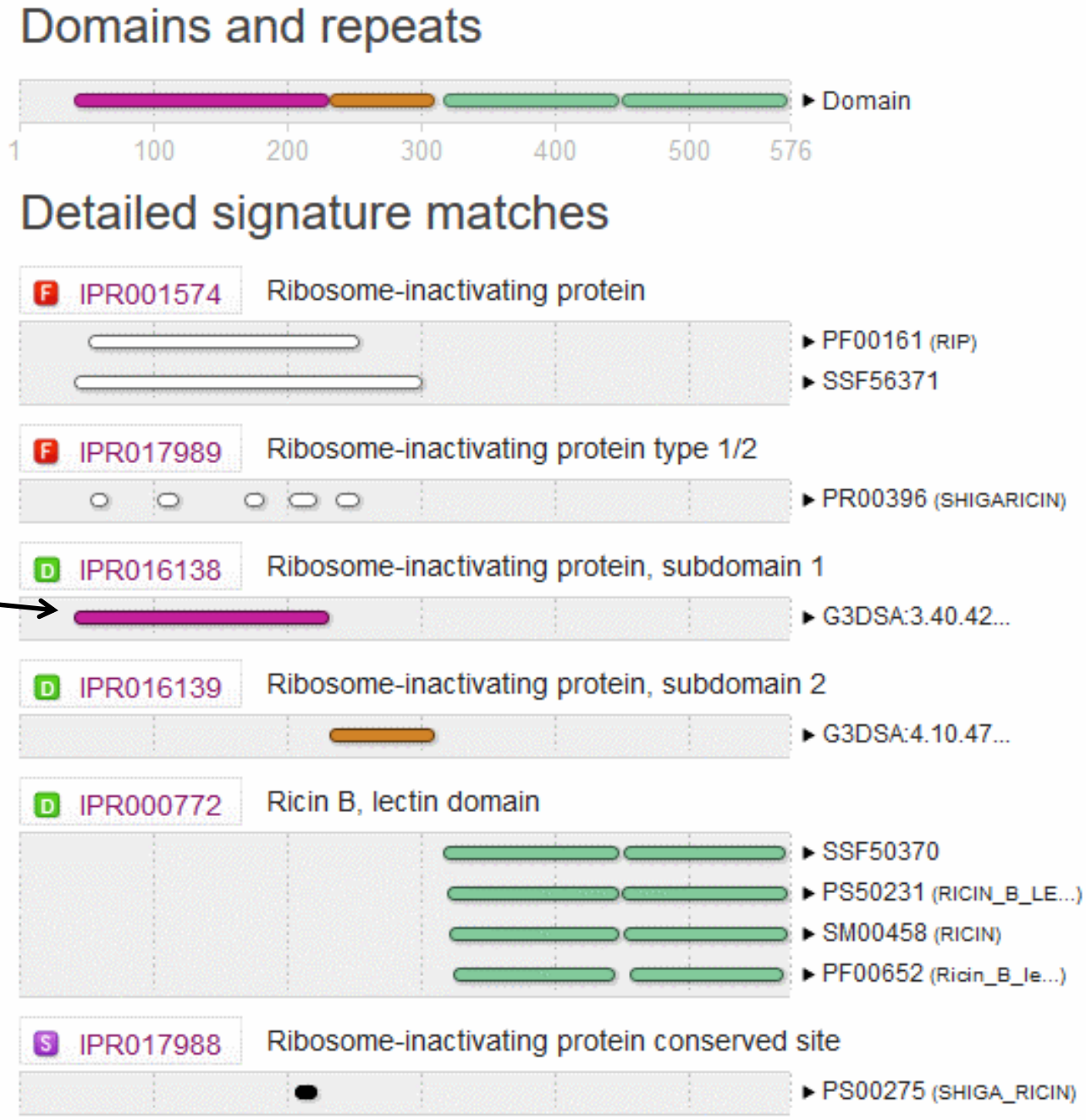
- RegEx Profiling: PROSITE
- PSSM fingerprints: PRINTS
- automated clustering: ProDom
- HMMs: Pfam, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, PANTHER, Gene3D

UniProtKB provides sequences, and InterProScan conducts searching.

Hunter et al, *Nucl. Acids Res.* (2009) 37: D211-D215

http://www.ebi.ac.uk/interpro/

Castor bean ricin report



Links to structures

Multiple names for one feature

http://www.ebi.ac.uk/interpro/protein/P02879

Takeaway messages

- Hidden Markov Models find likely protein-coding genes in assembled contigs.
- Aligning sequences was the original "killer app" of bioinformatics, and it continues to be critically important to molecular biology.
- Sequence similarity often implies evolutionary relationship. It often implies related structure and function.