Sequences, Structures, and Gene Regulatory Networks

Learning Outcomes

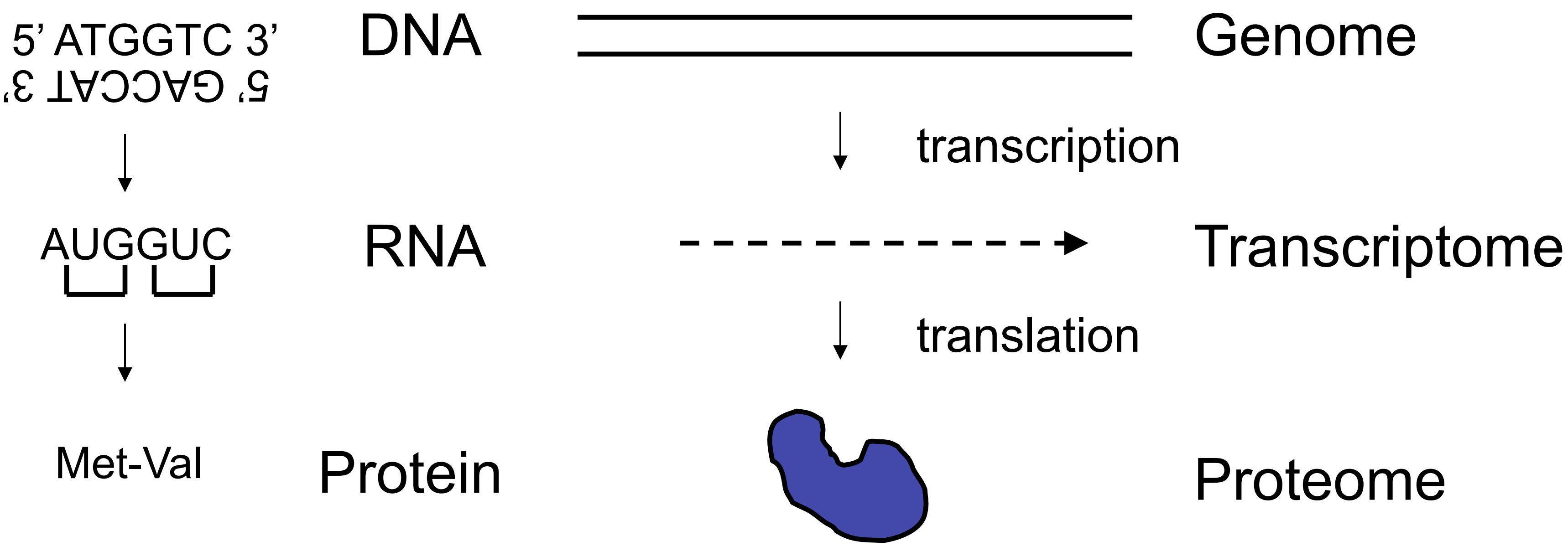
After this class, you will

- Understand gene expression and protein structure in more detail
- Appreciate why biologists like to align sequences, and have a general idea of how the most commonly used algorithm, BLAST, works
- Be able to use your knowledge of biology to help you critique visual representations of alignments and gene regulatory networks

Outline

- Sequences and Structures: The Central Dogma in a little more detail
- Alignment – why is this so important? What are important features to visualize?
- Gene regulatory networks
- Appendix: Representation of sequences in databases at the NCBI

The Central Dogma of Molecular Biology: Genes Encode Proteins



5’ ATGGTC 3’
5’ GACCAT 3’

AUGGUC

Met-Val

DNA

RNA

Protein

transcription

translation

Genome

Transcriptome

Proteome

Prokaryotic and eukaryotic transcription and translation

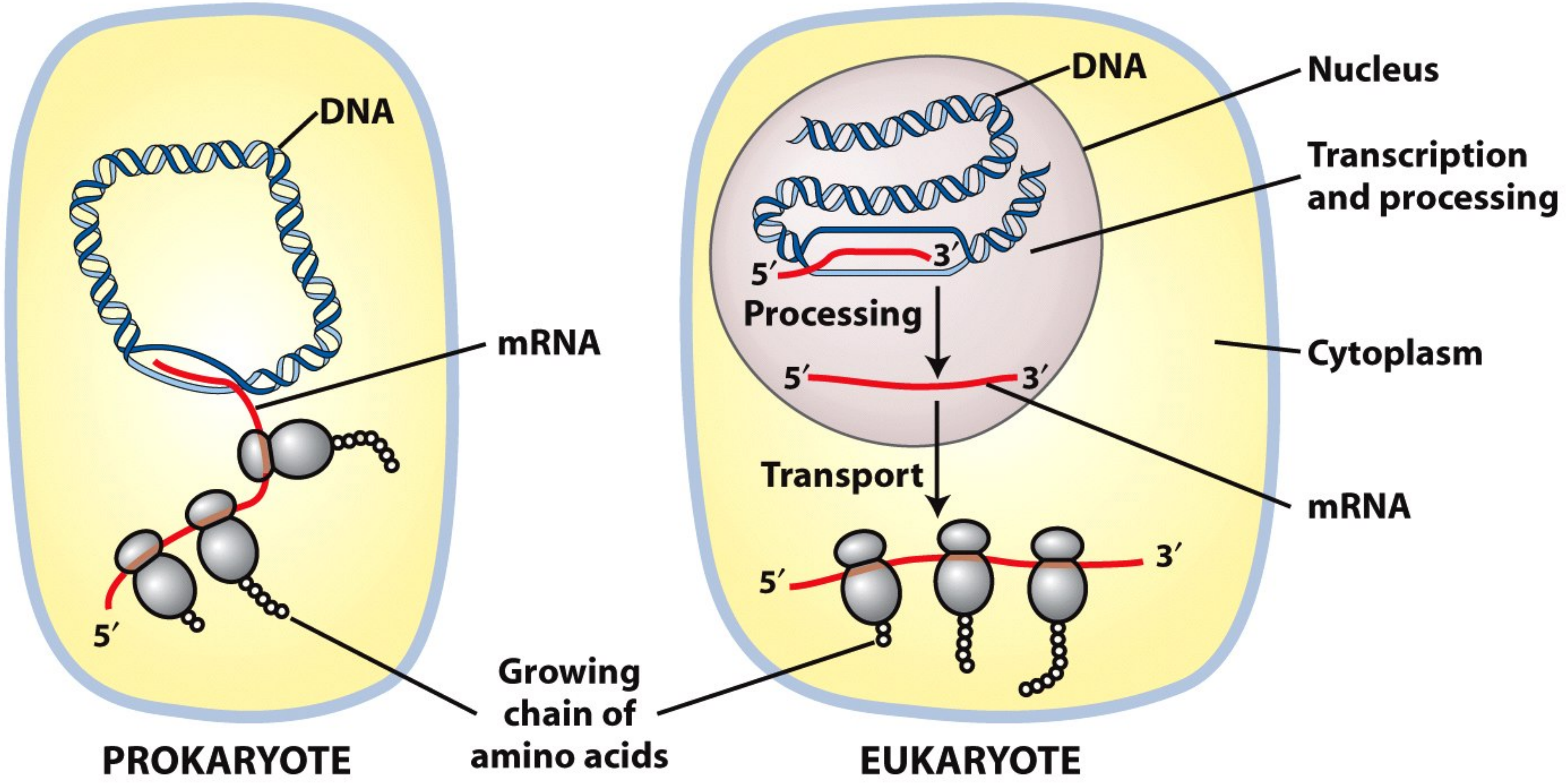
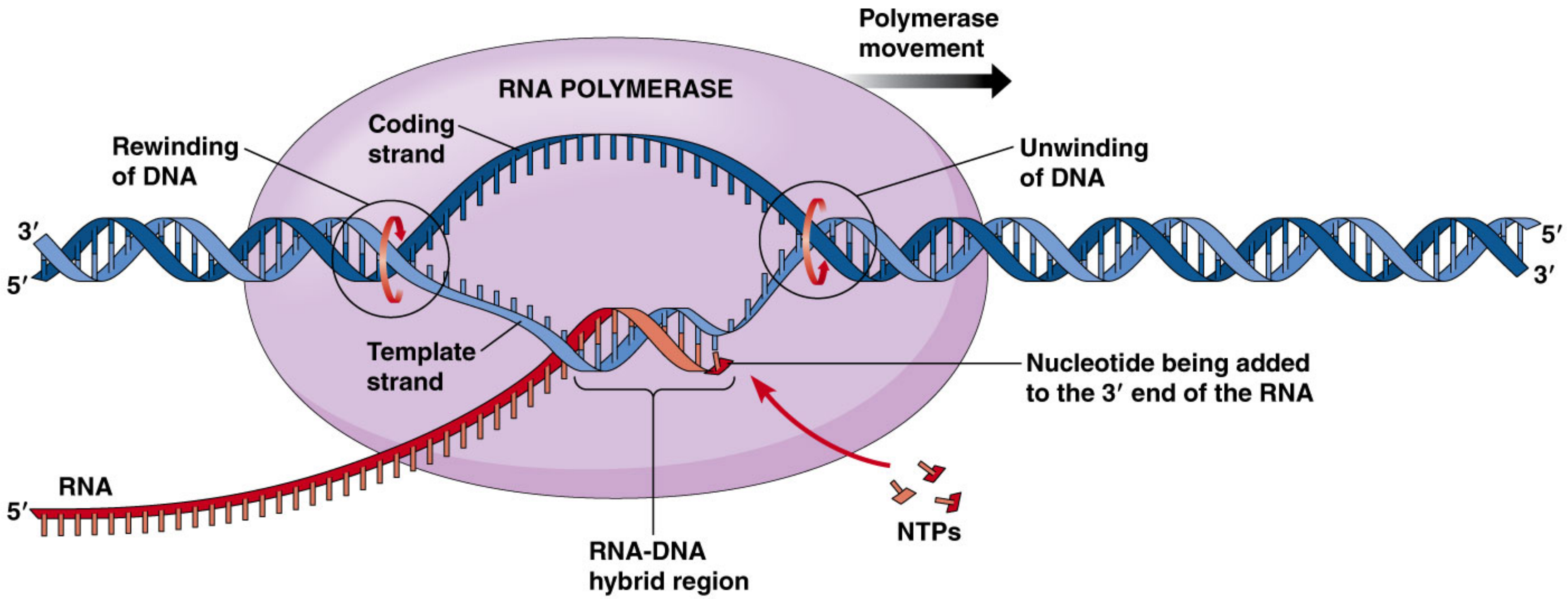


Figure 8-11
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

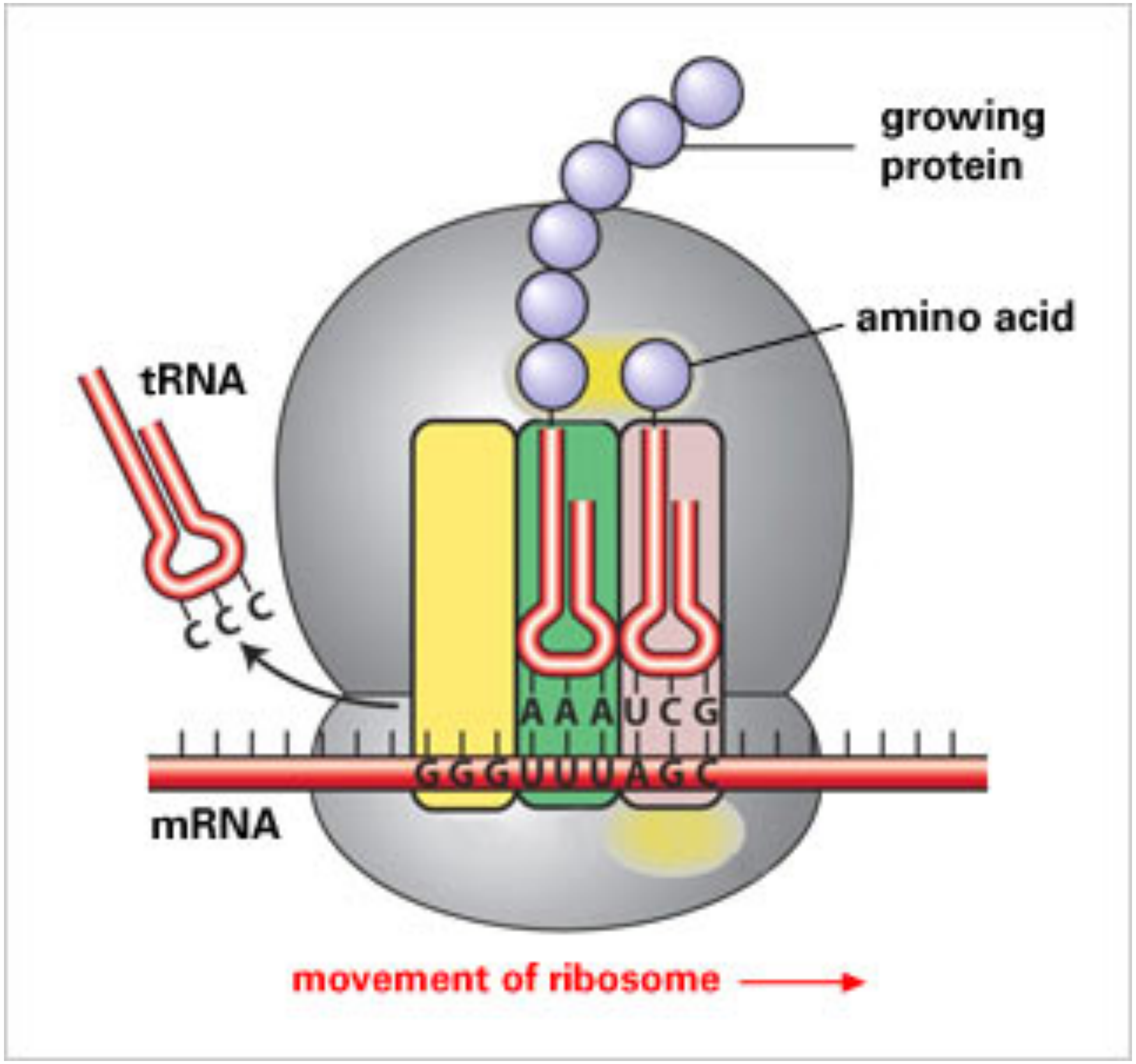
Griffeths, Introduction to Genetic Analysis, 2008

Transcription is mediated by RNA Polymerase



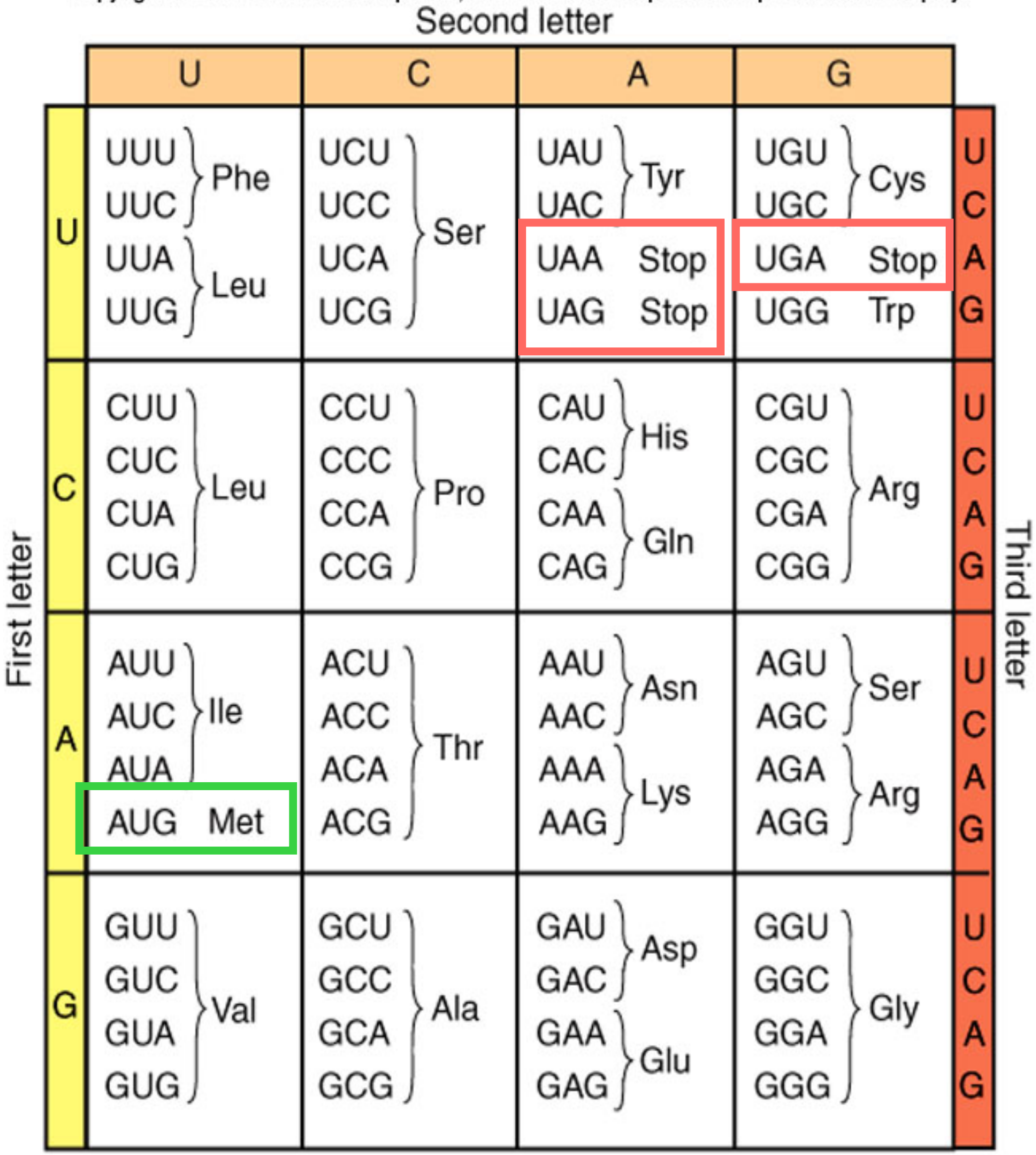
© 2012 Pearson Education, Inc.

Translation is mediated by ribosomes



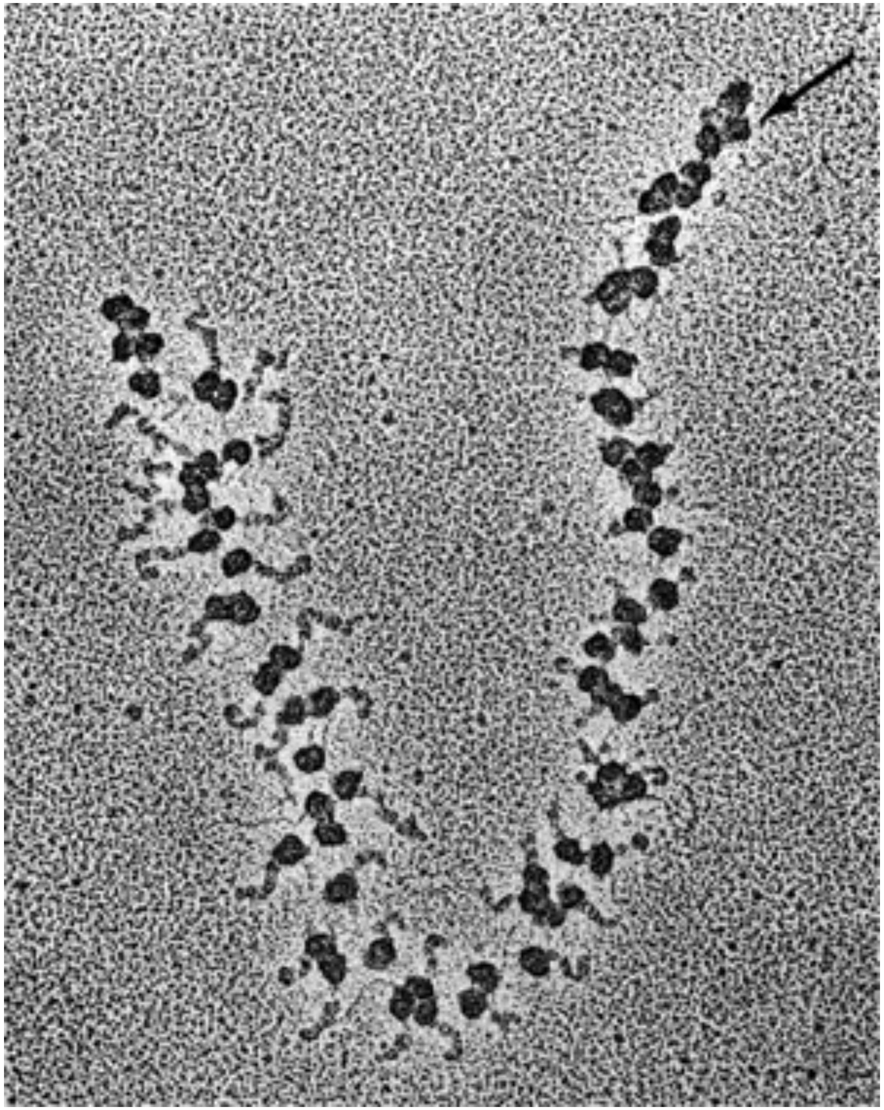
http://www.tokresource.org/tok_classes/biobiobio/biomenu/transcription_translation/

The Genetic Code



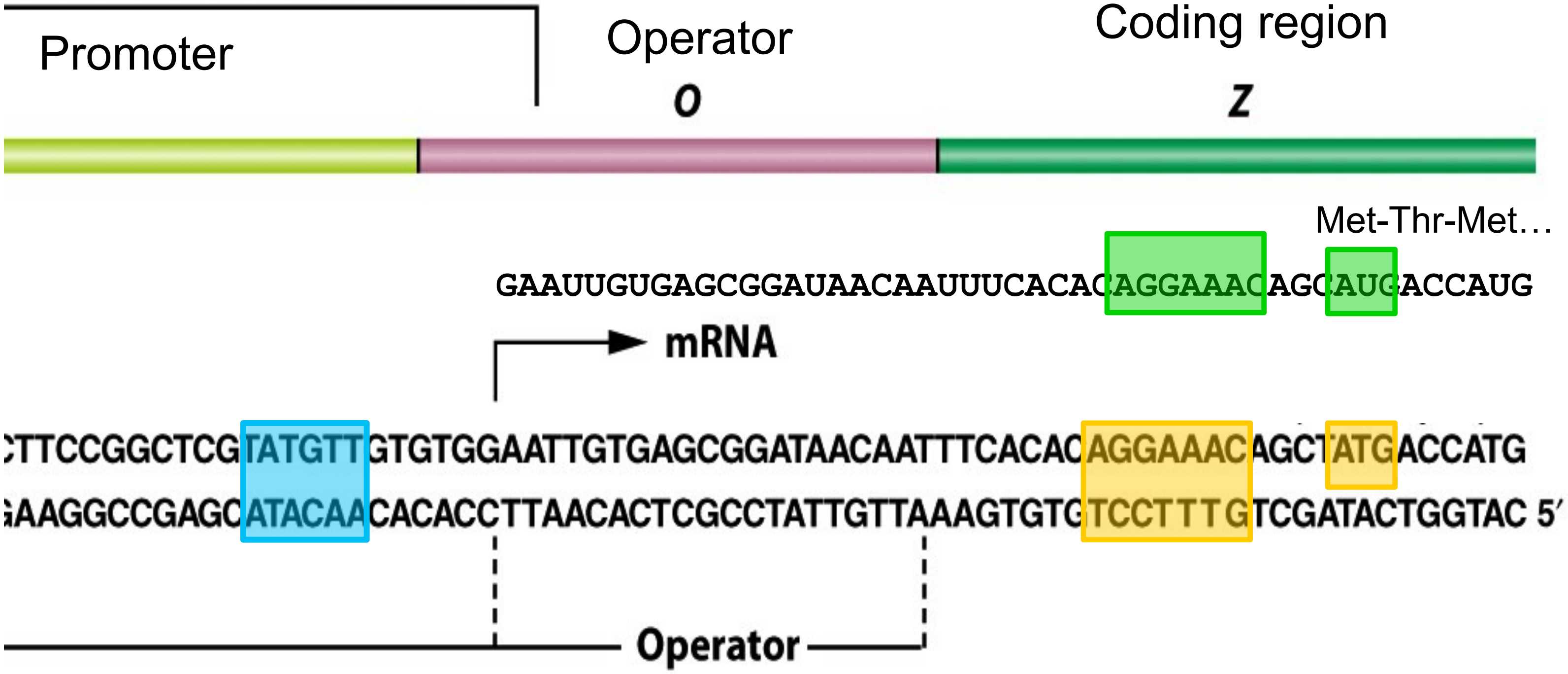
First letter	Second letter								Third letter
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Some evolutionary thoughts



http://courses.bio.indiana.edu/L104-Bonner/F09/imagesF09/L23/Ribosomes.html

Transcription vs. Translation: Lac operon control region


Promoter
Operator
O
Coding region
Z
Met-Thr-Met…
GAAUUGUGAGCGGAUAACAAUUUCACACAGGAAACAGCUAUGACCAUG
mRNA
CTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATG
GAAGGCCGAGCATACAACACACCTTAACACTCGCCTATTGTTAAAGTGTGTCCTTTGTCGATACTGGTAC 5′
Operator

alysis, Ninth Edition
ompany

Signals are not perfect matches to consensus sequences

🟦 RNAP recognizes sequence in promoter DNA

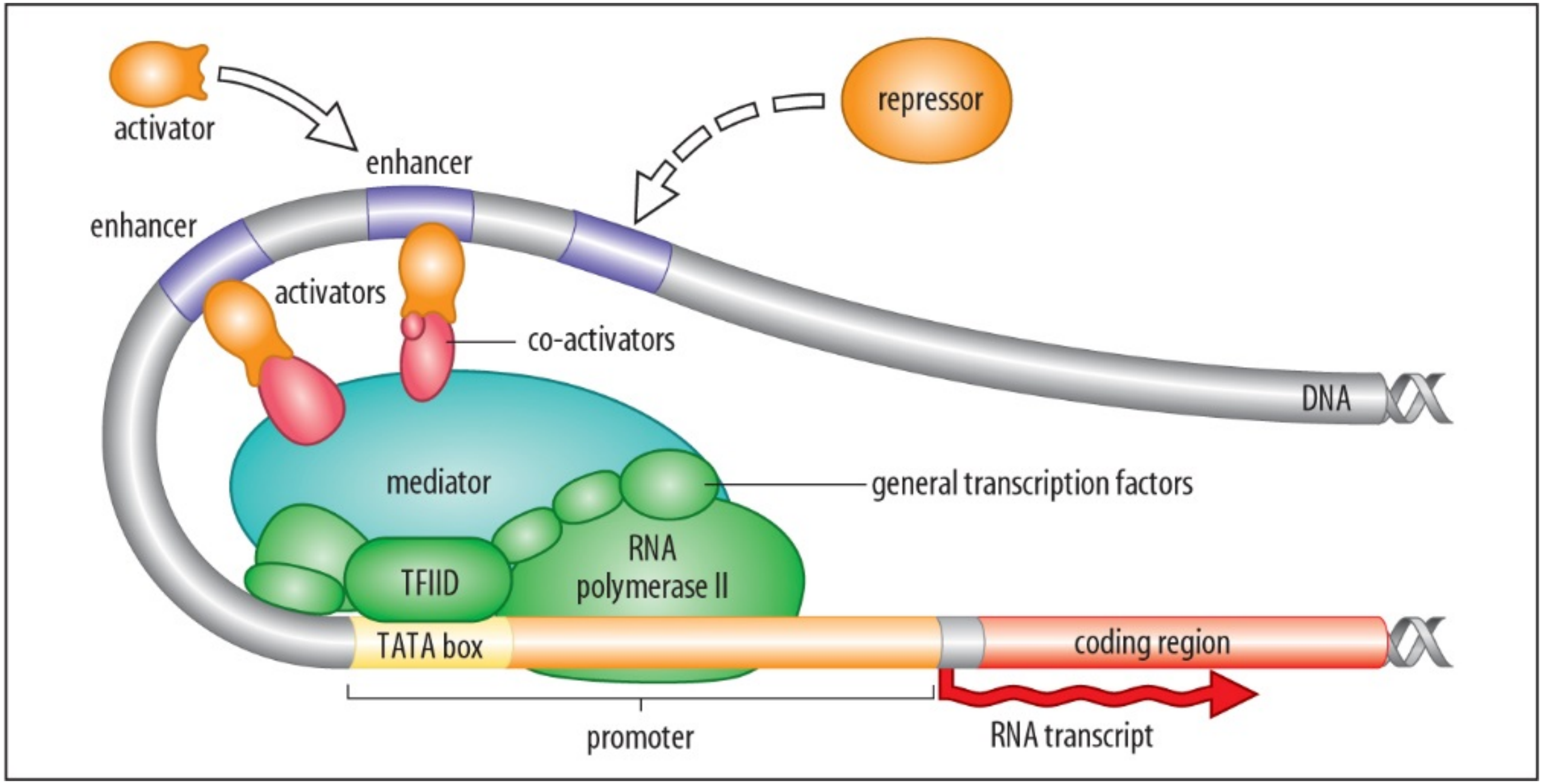
🟩 Ribosome recognizes signal in RNA

🟨 Seen by computer in DNA

RNAP does not care about codons; it cannot distinguish coding and non-coding DNA.

1º transcript == mRNA in bacteria; no splicing, capping, polyA

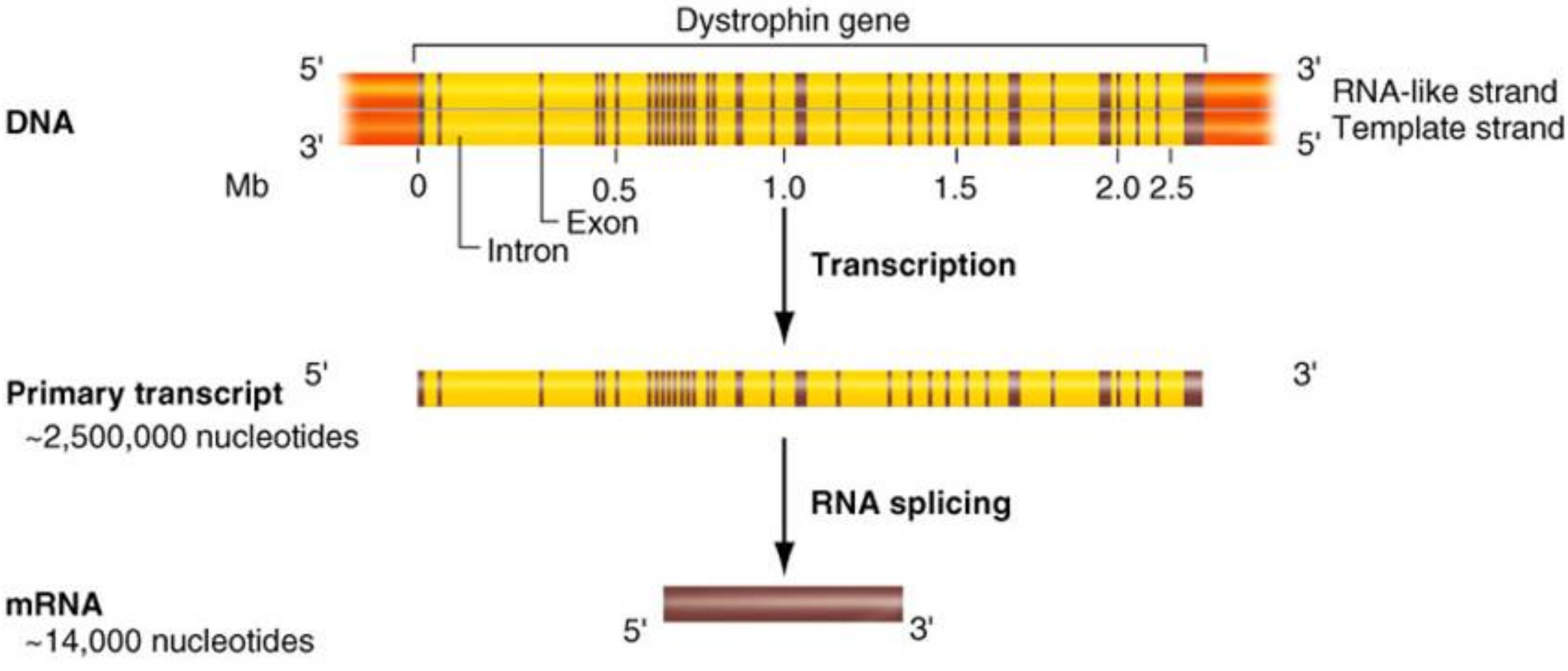
Eukaryotic transcription is complex!



- Basal transcriptional regulators
- Cell type specific enhancers and repressors

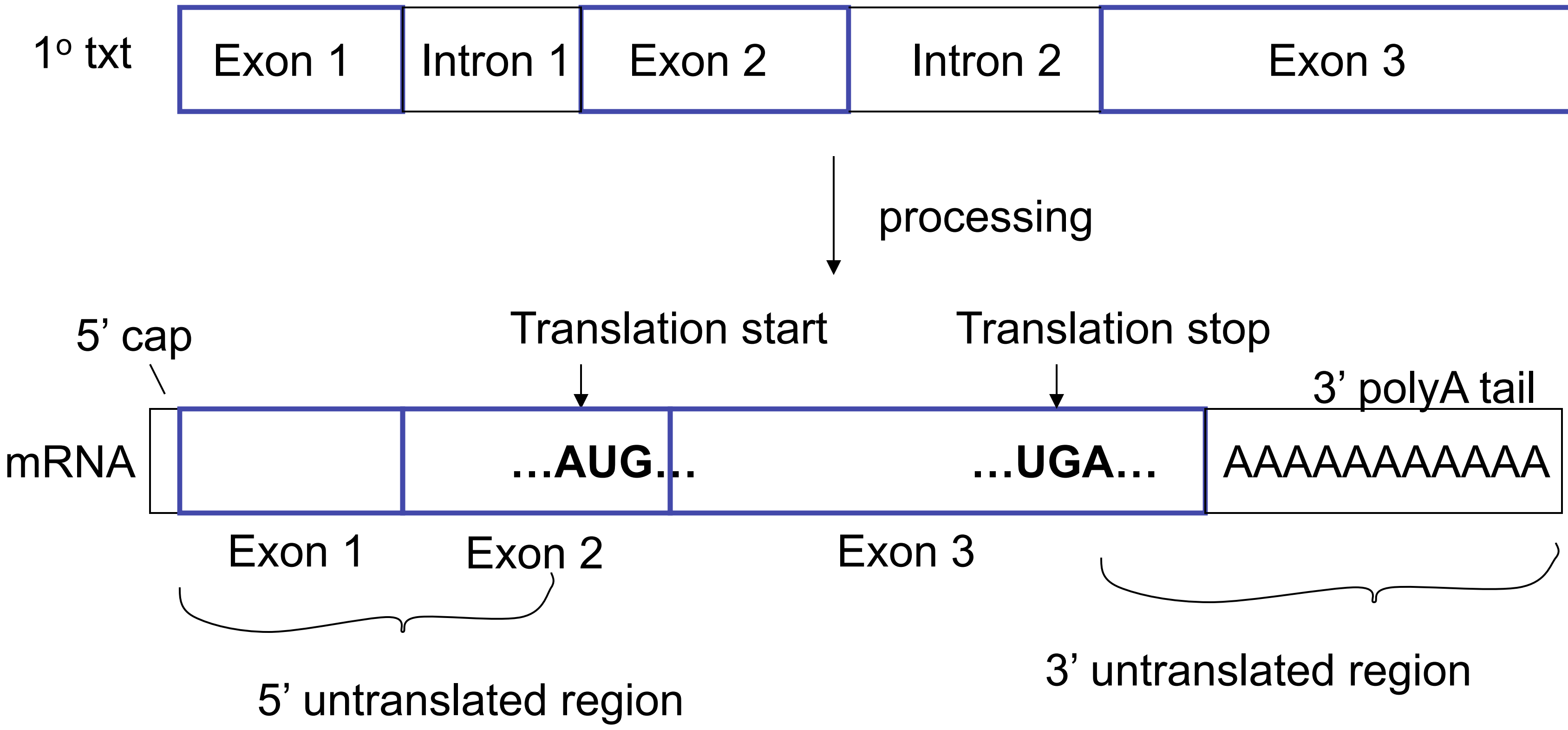
http://www.mun.ca/biology/desmid/brian/BIOL3530/DEVO_10/devo_10.html

RNA processing in more complex eukaryotes



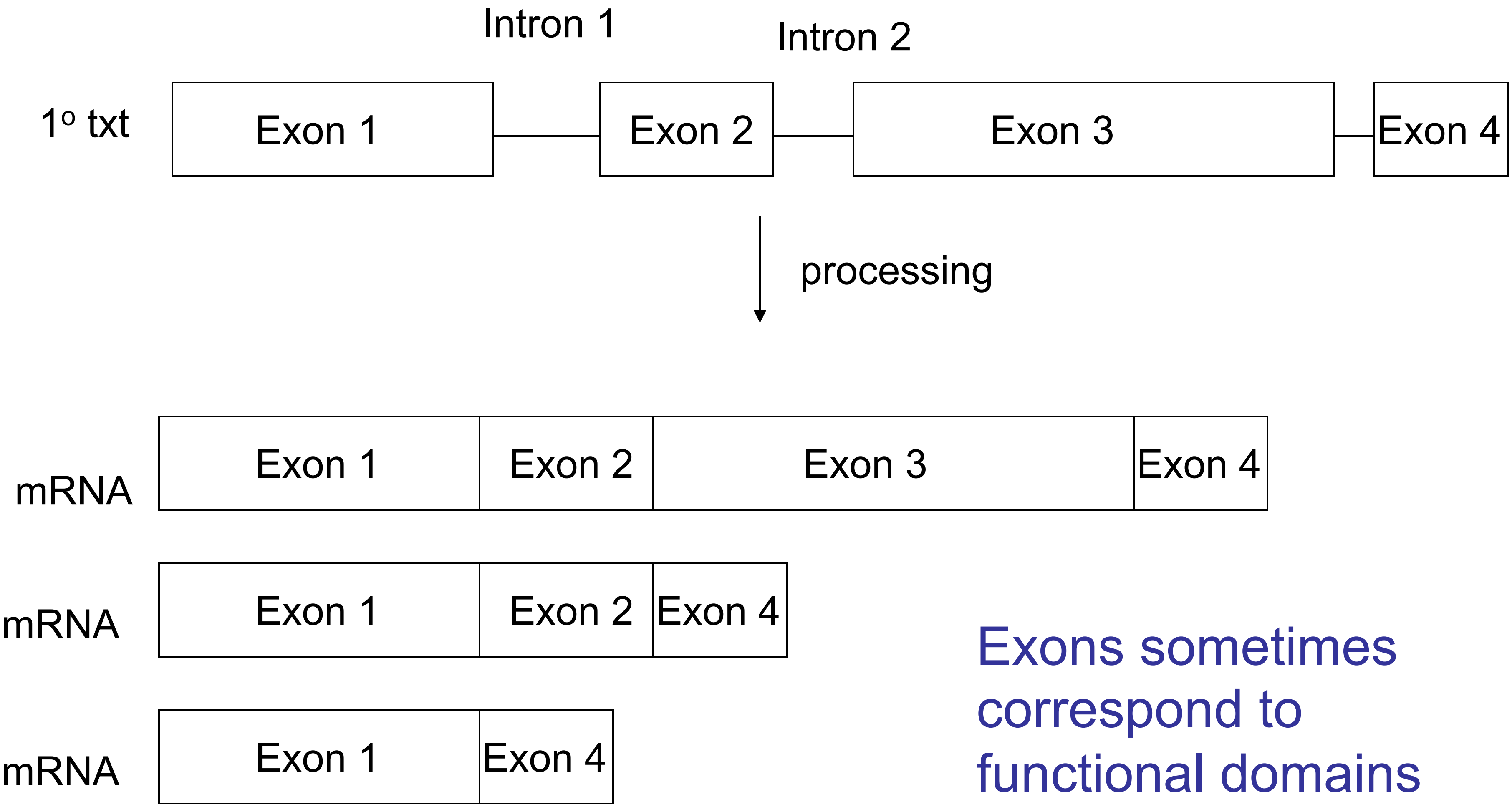
Hartwell et al. Genetics: From Genes to Genomes

Schematic view of processing of mRNA in eukaryotes



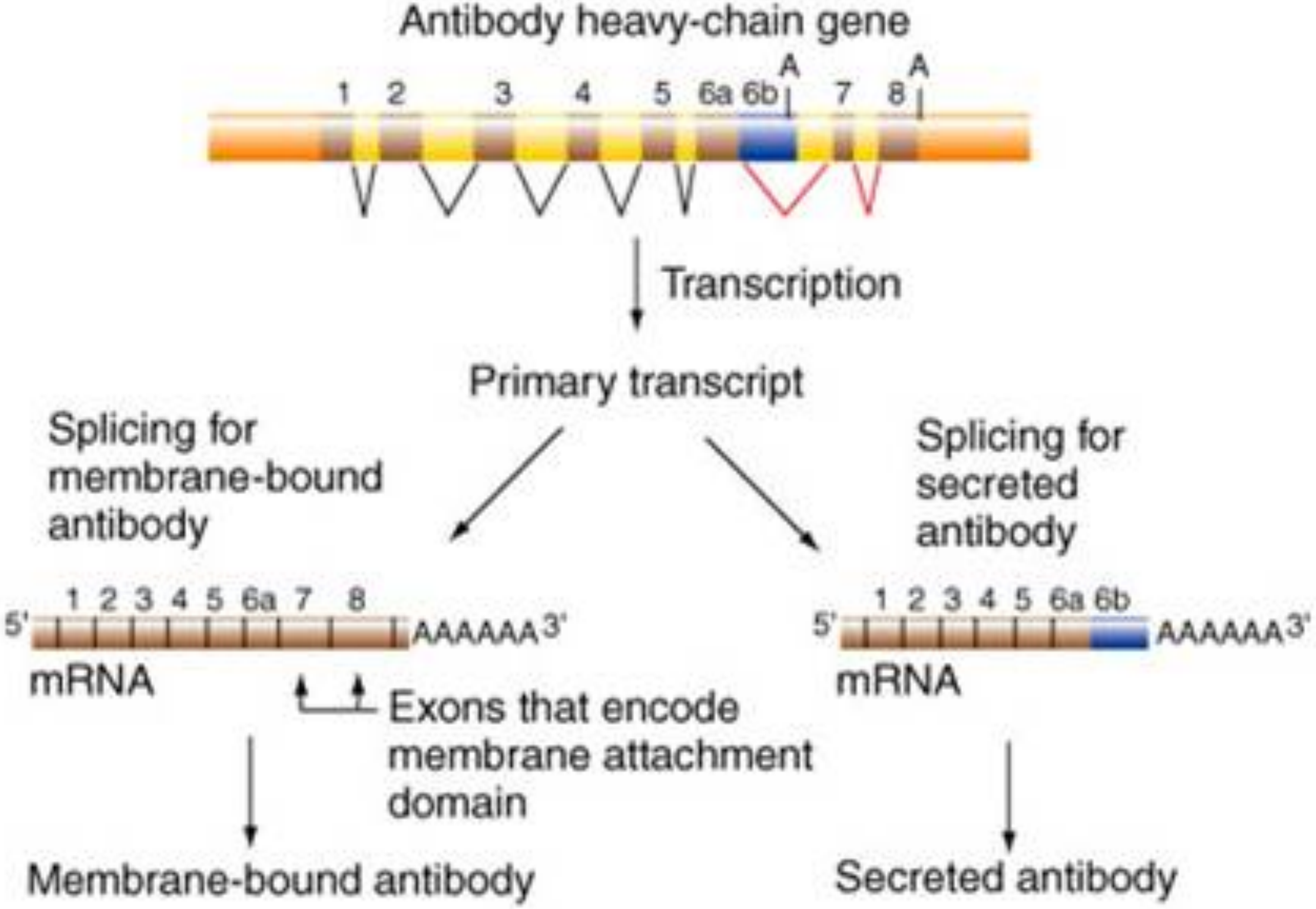
Is the entire transcript translated?

Alternative splicing



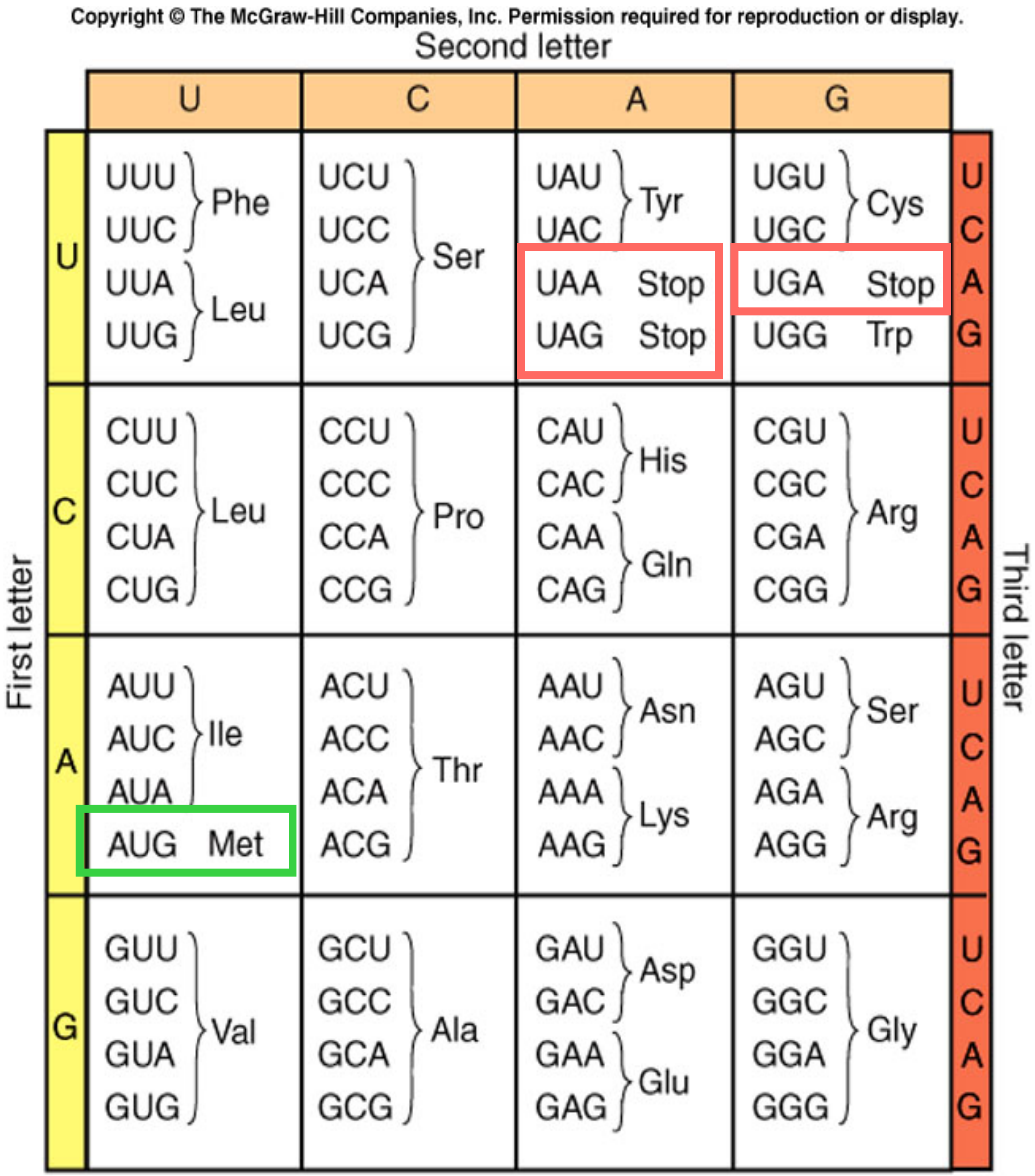
Exons sometimes correspond to functional domains of proteins

Different splice forms can function very differently in the cell



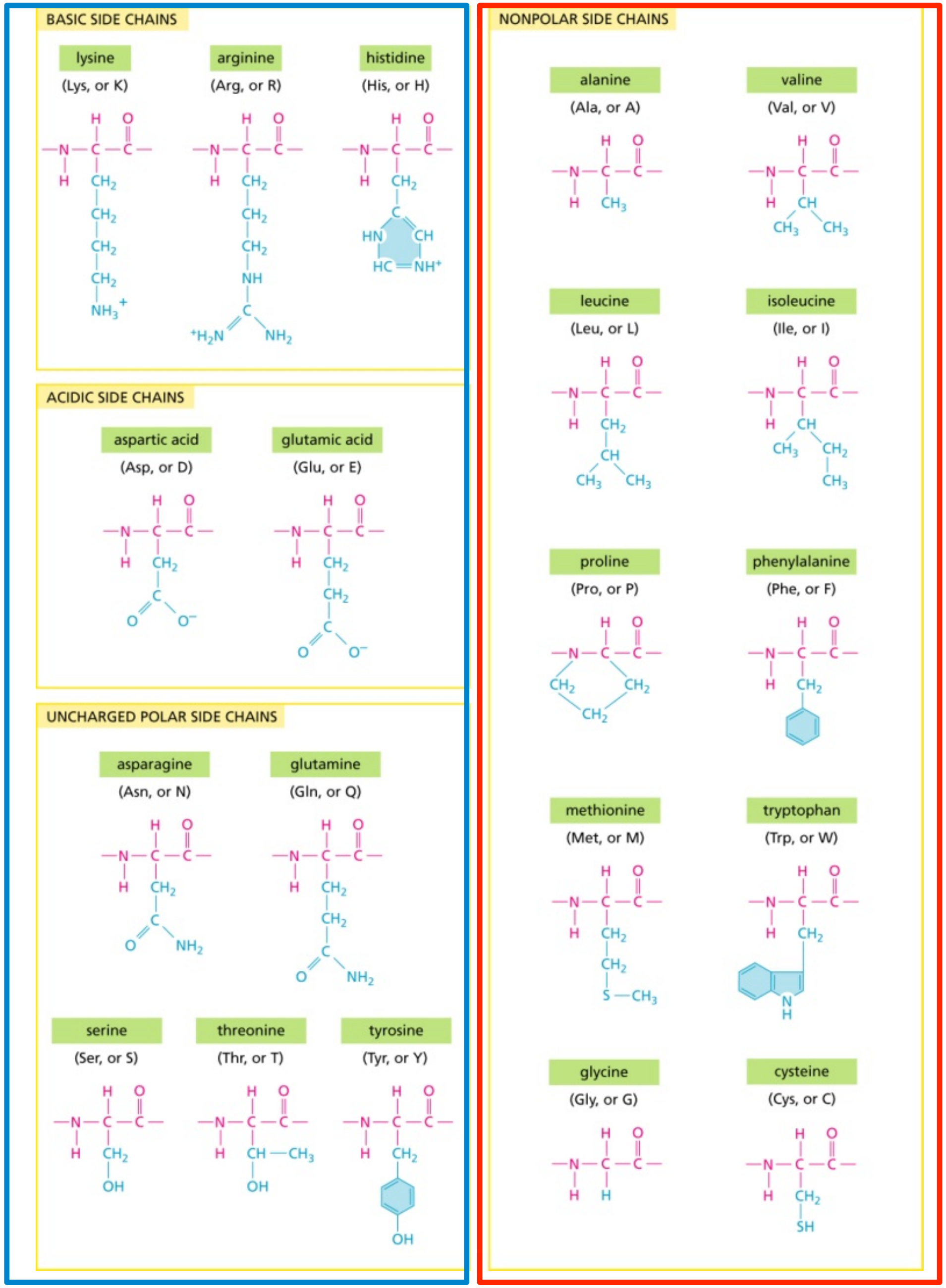
Hartwell et al. Genetics: From Genes to Genomes

The Genetic Code



Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA Stop	UGA Stop	A
	UUG	UCG	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lys	AGA Arg	A
	AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glu	GGA	A
	GUG	GCG	GAG	GGG	G

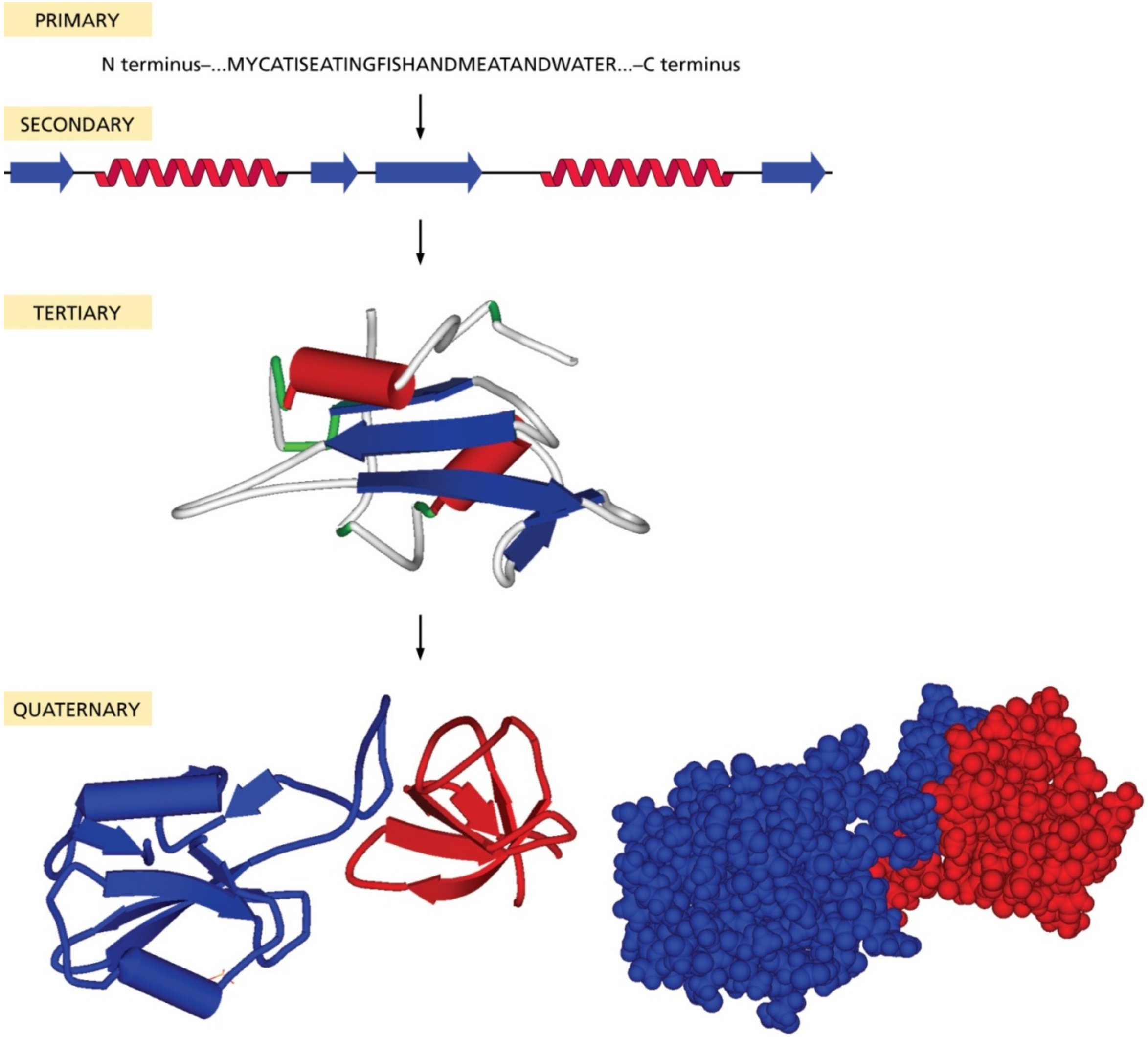


Amino acids fall into different classes

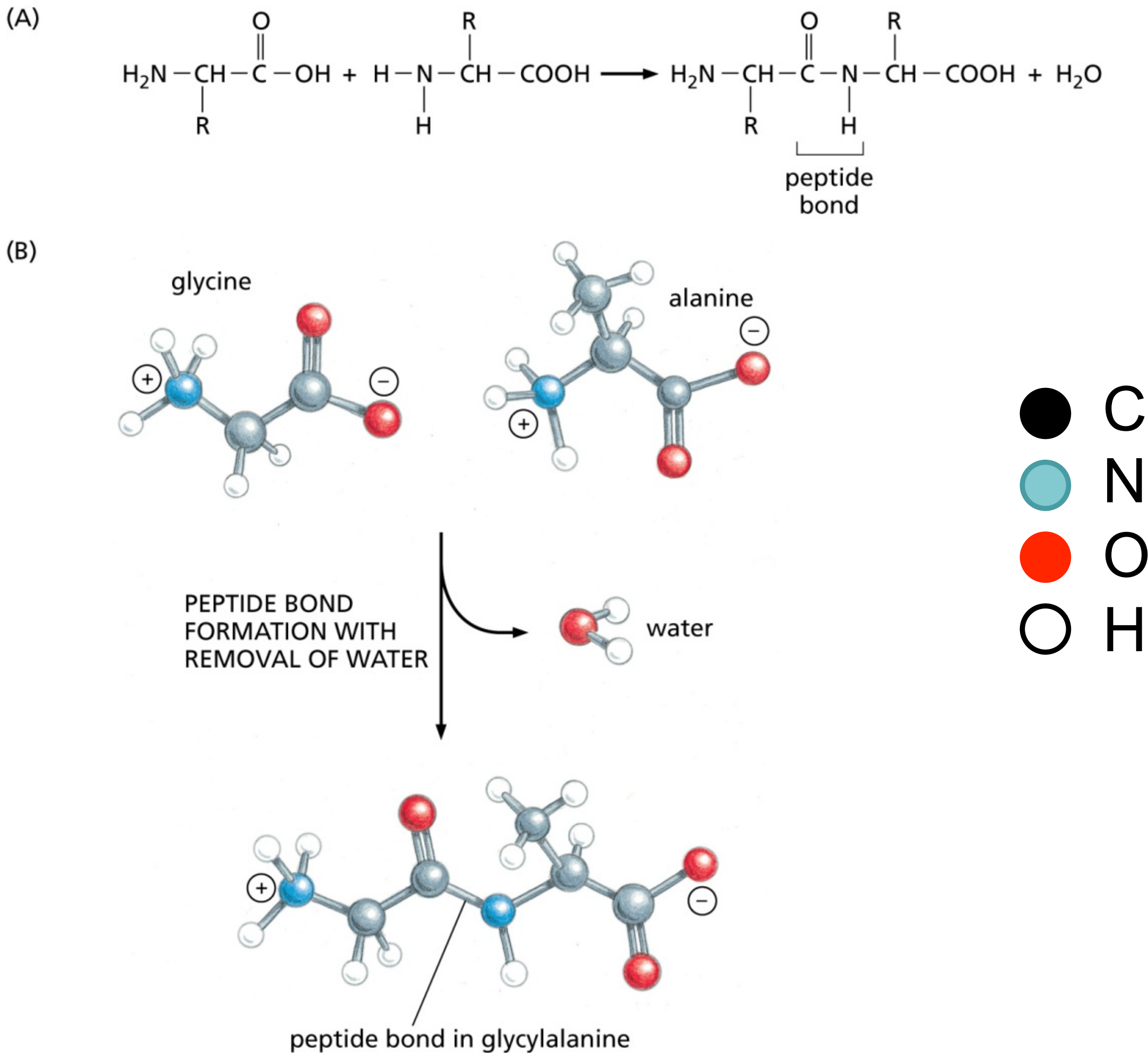
Hydrophobic:
Nonpolar side chains often found in protein core, transmembrane regions

Hydrophilic:
Polar and charged side chains often found nearer protein surface, interact with water

Four levels of protein structure


PRIMARY
N terminus–...MYCATISEATINGFISHANDMEATANDWATER...–C terminus
SECONDARY
TERTIARY
QUATERNARY

Making a polypeptide chain: primary structure


(A)
peptide bond
(B)
glycine
alanine
PEPTIDE BOND FORMATION WITH REMOVAL OF WATER
water
peptide bond in glycylalanine

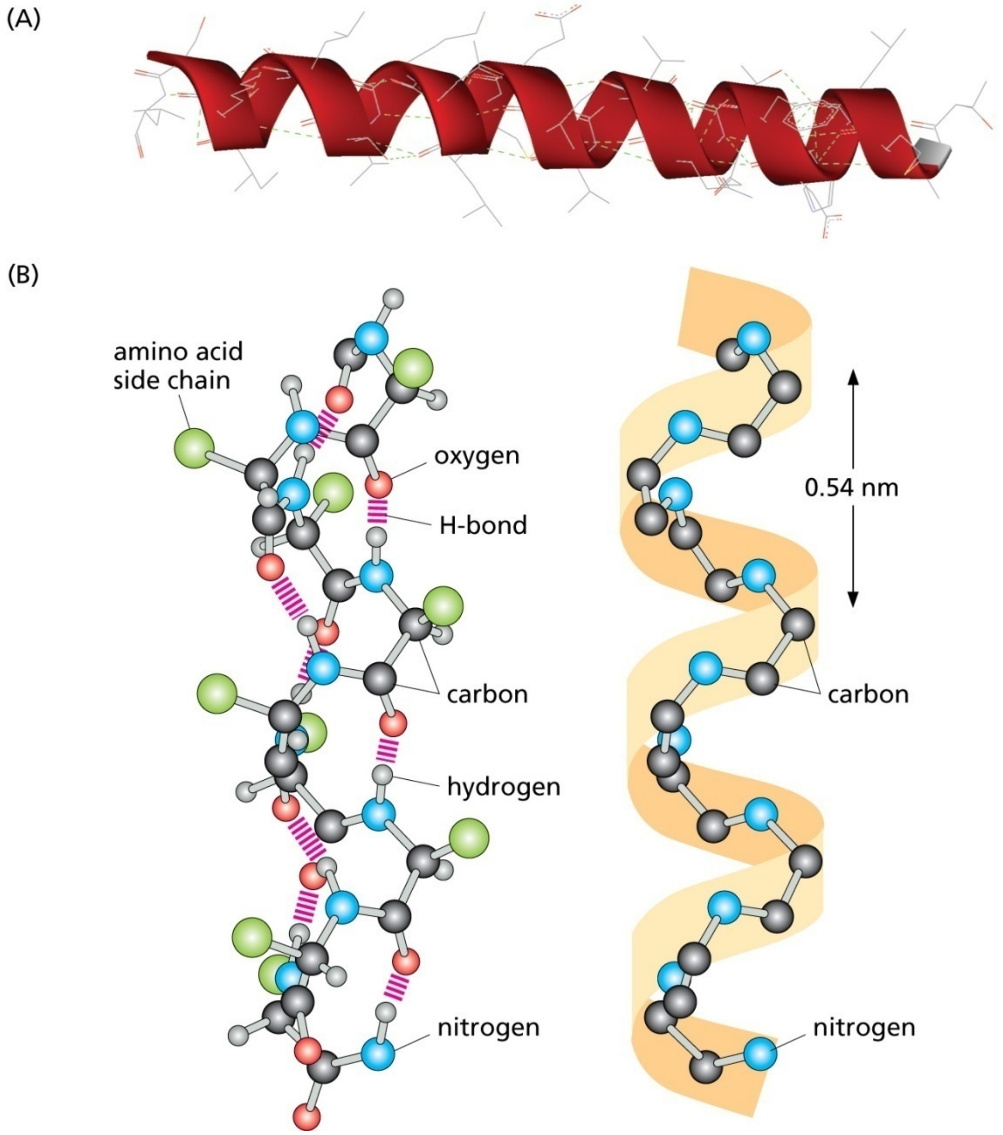
● Carbon

● Nitrogen

● Oxygen

○ Hydrogen

Secondary structure: α-helices



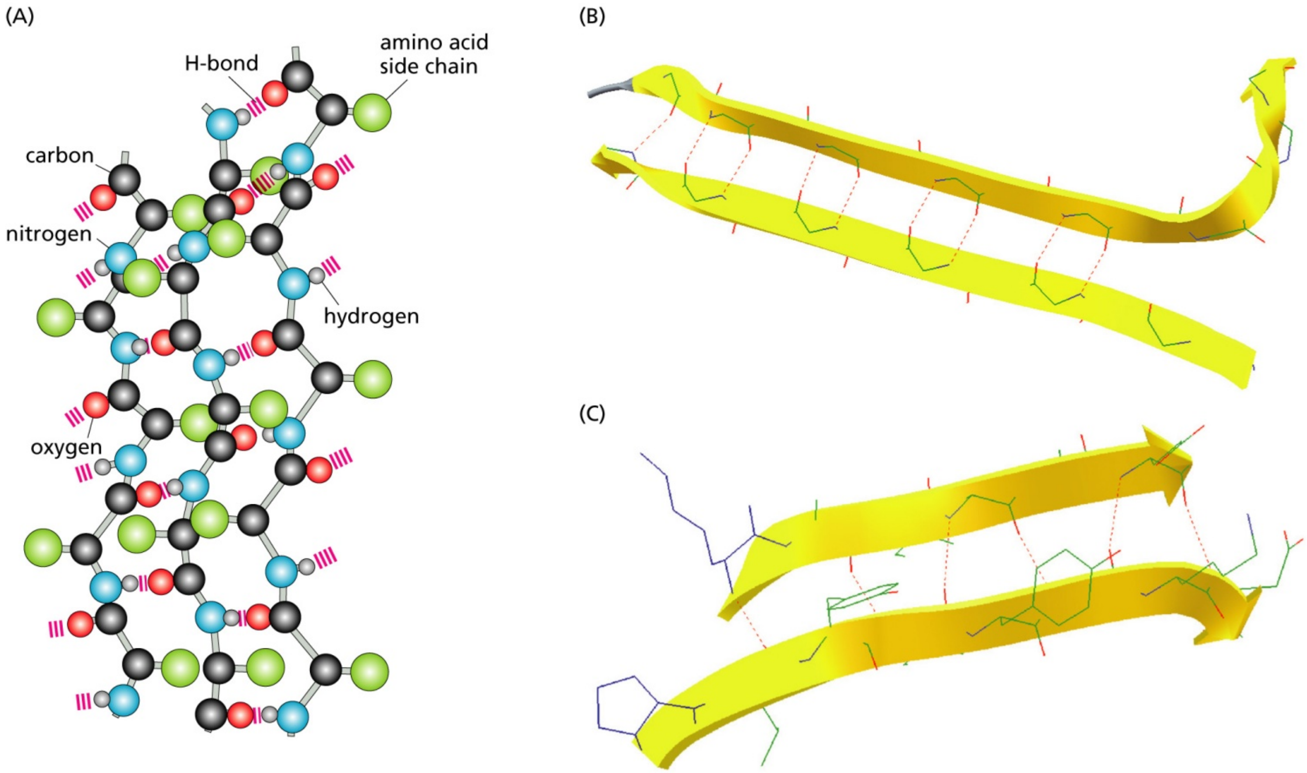
Ala, Glu, Leu, Met 'like' α-helices

Pro rarely found in helices

Gly, Tyr, also poor helix formers

● Carbon
● Nitrogen
● Oxygen
○ Hydrogen

Secondary structure: β-strands



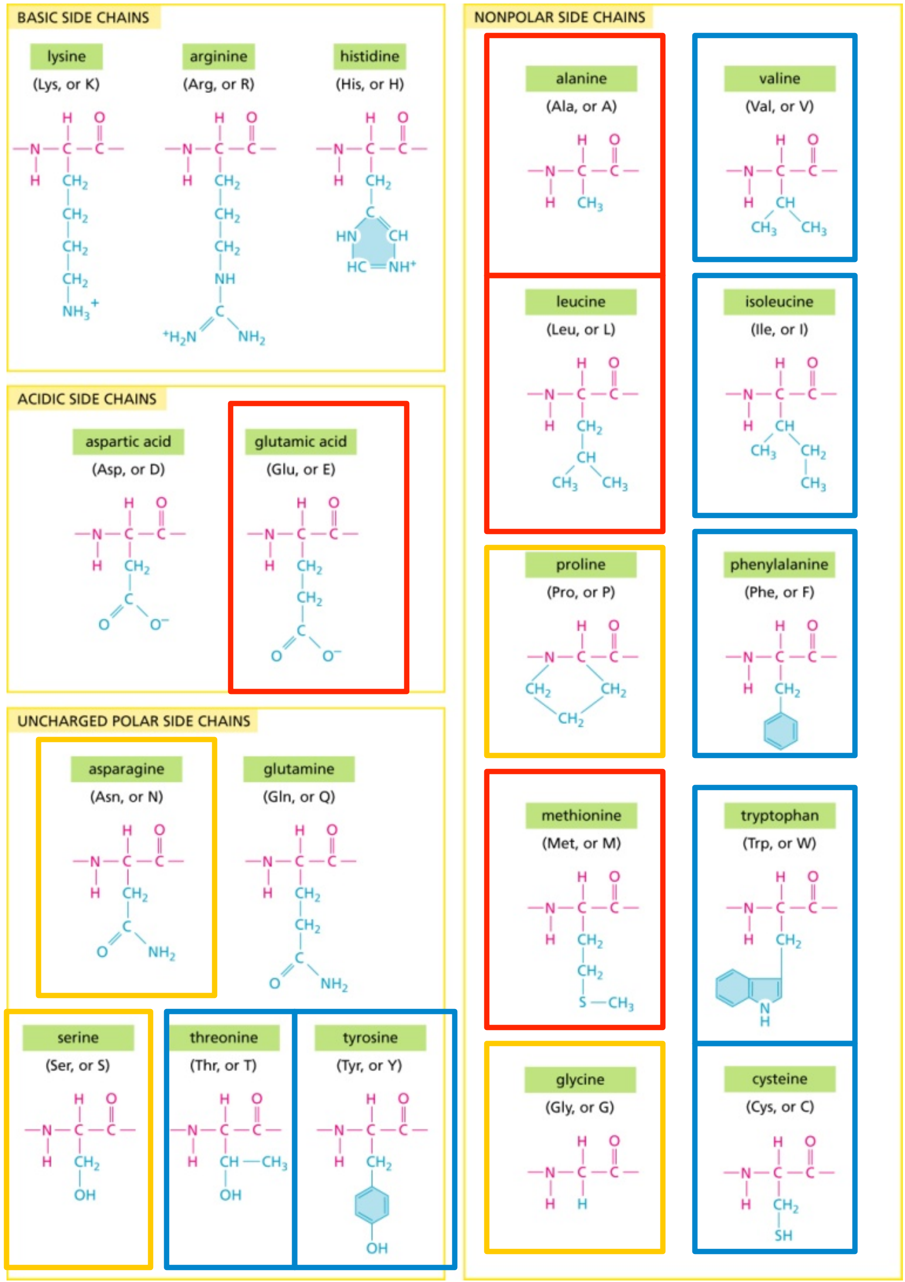
● Carbon

🔵 Nitrogen

🔴 Oxygen

○ Hydrogen

Val, Ile, Tyr, Cys, Trp, Phe, Thr ‘like’ β-strands



Amino acids fall into different classes

🟥 α-helix formers

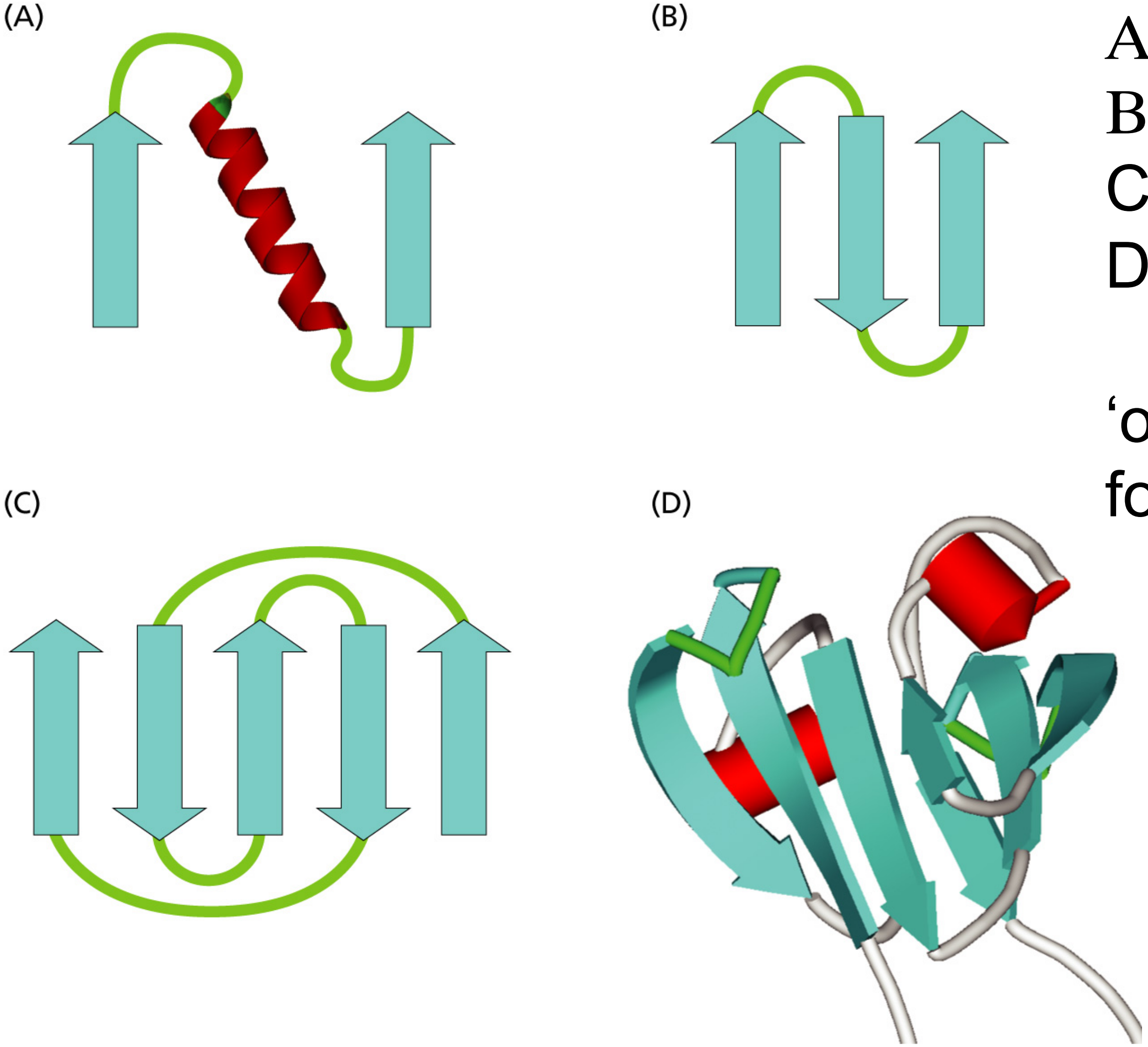
🟦 β-strand formers

🟨 Turn segments

Gly, Tyr, and especially Pro are poor α-helix formers

Some evolutionary thoughts

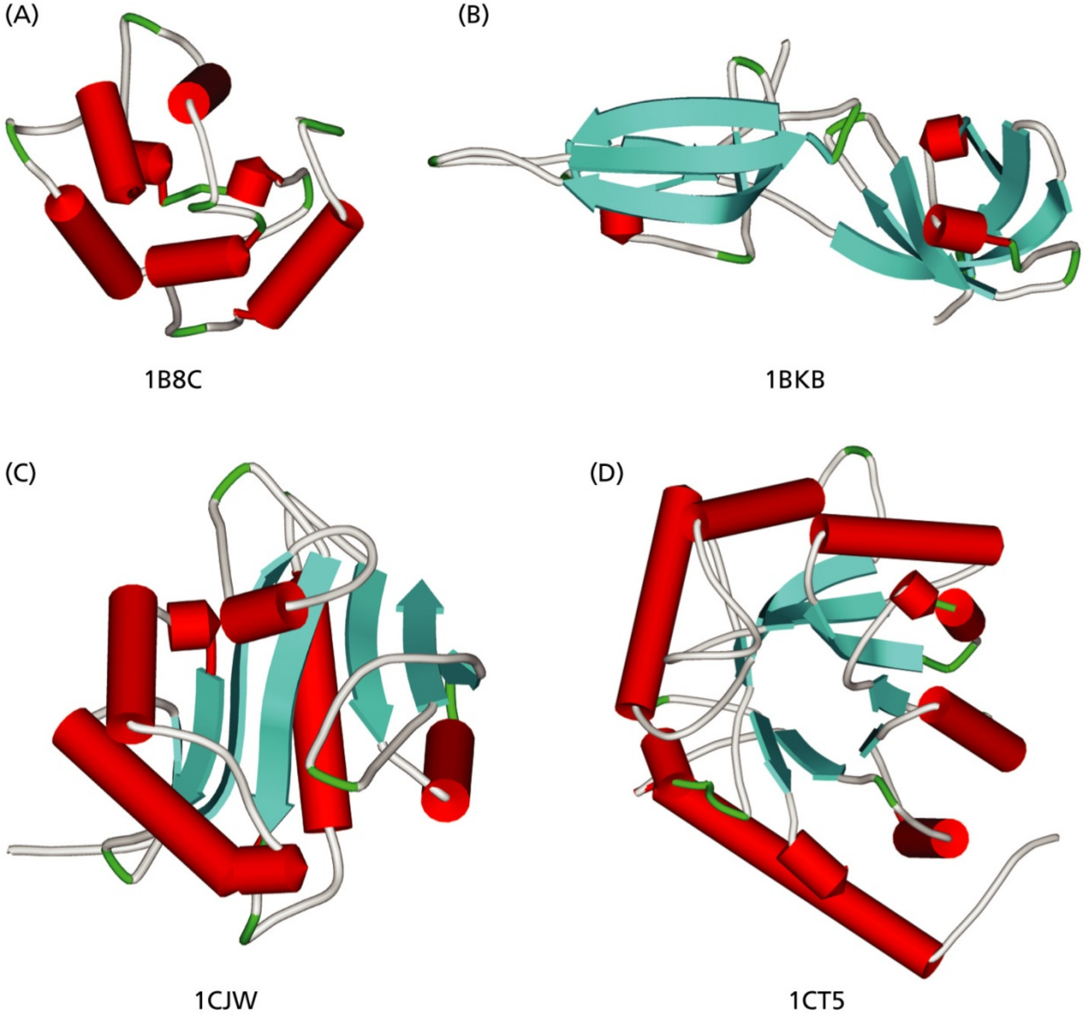
Supersecondary structure



A) βαβ repeat
B) β meander
C) Greek Key
D) Greek Key, β-crystallin

‘only’ 2,000 fold families for 35,000 structures

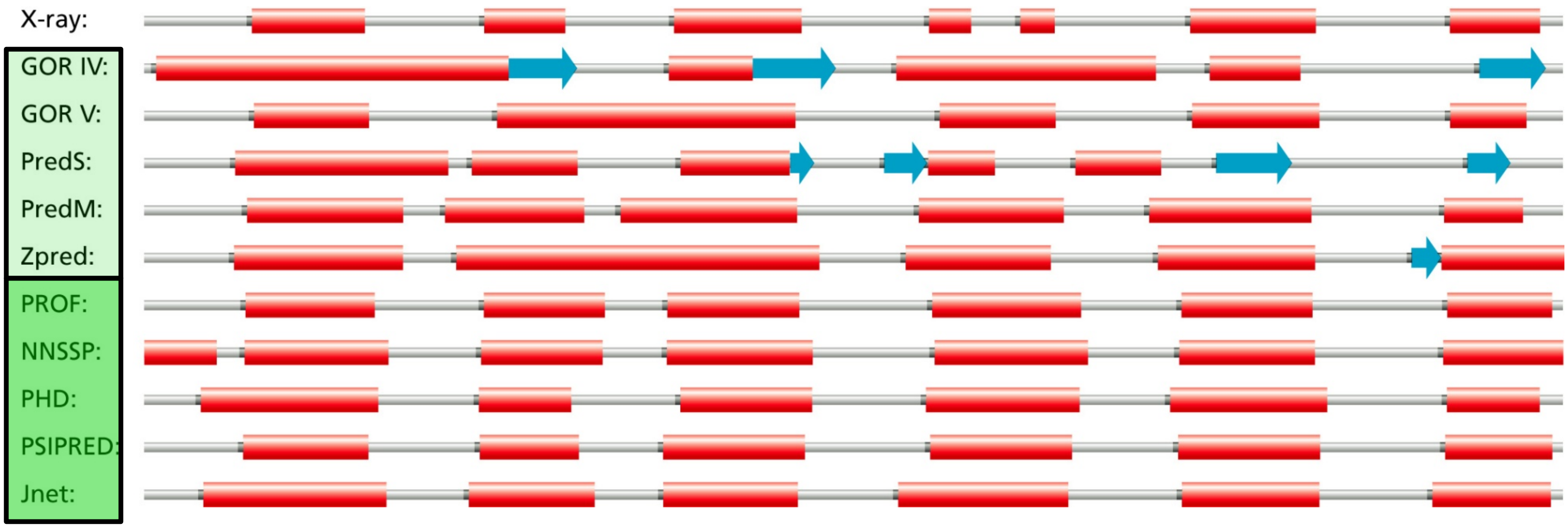
Prediction Examples: Known Structures

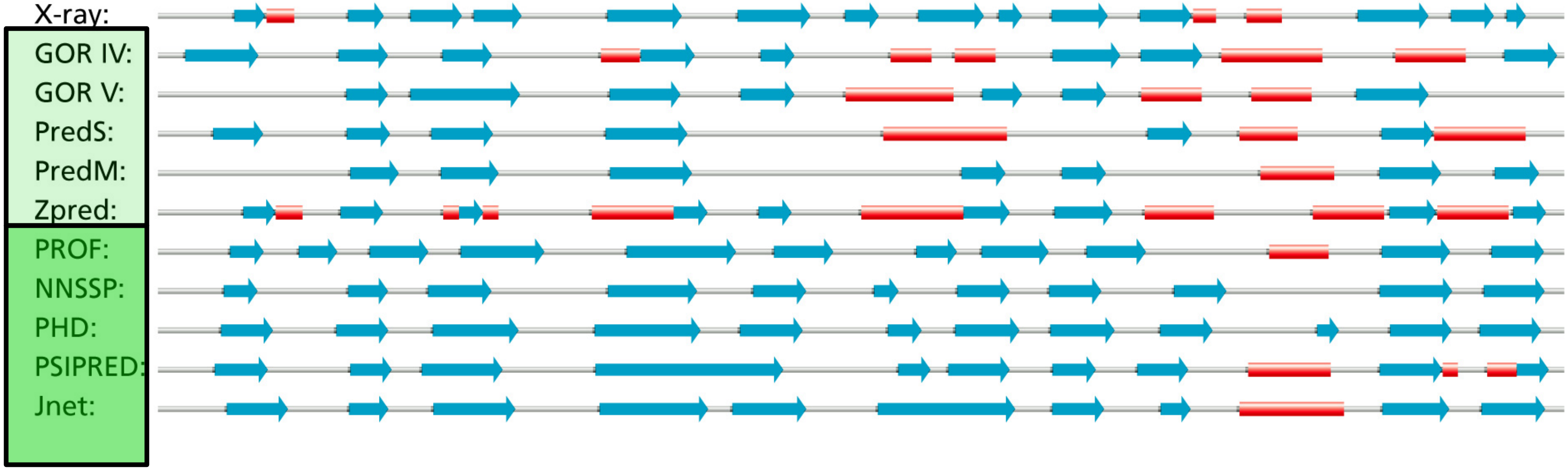


Knowledge based

α helix / β sheet

Neural net

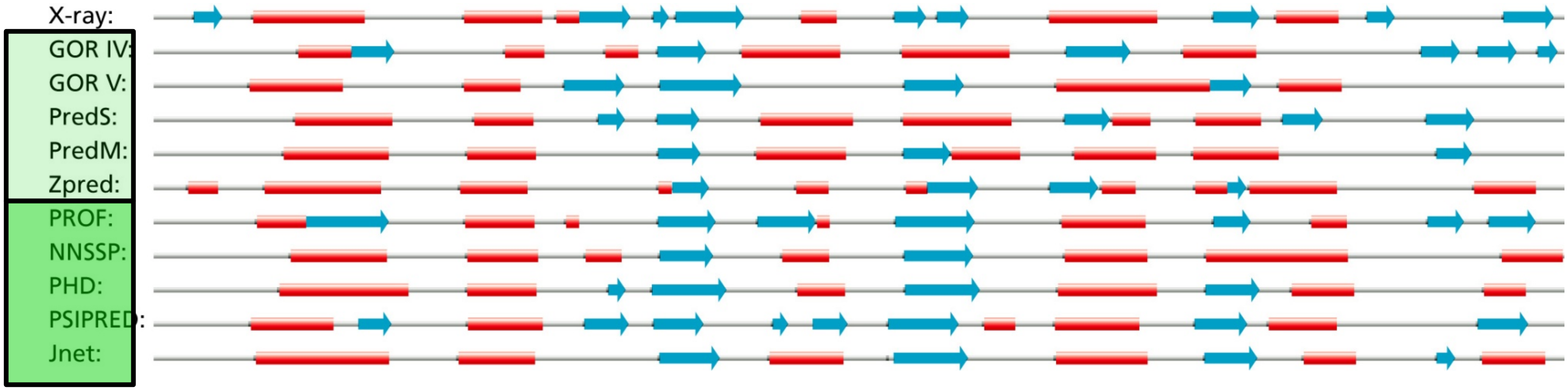


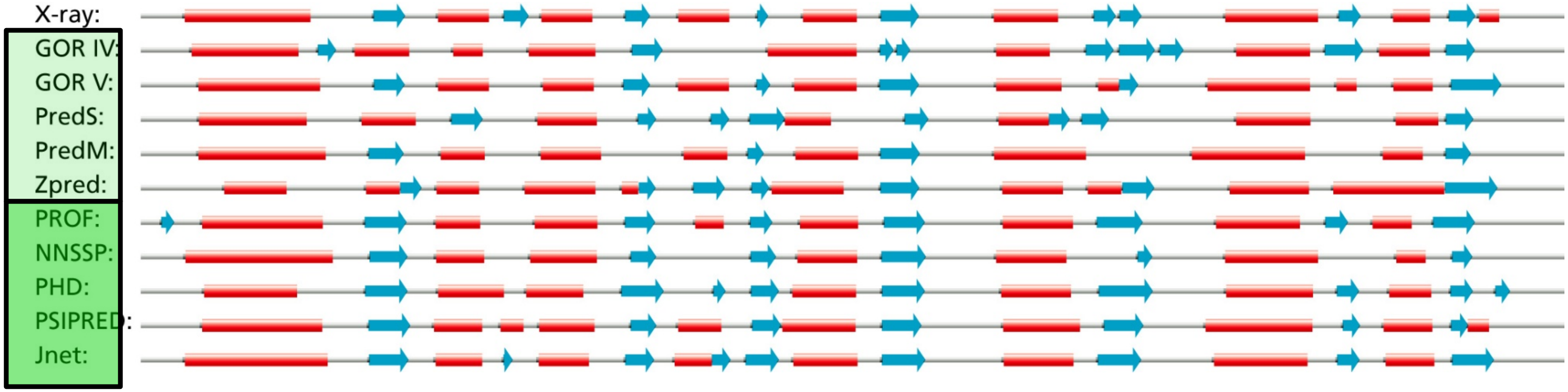


α + β and α / β fold

Knowledge based

Neural net





Some evolutionary thoughts

- Mutations occur at random in ___________
- Are all mutations bad? The Genetic Code
- Are some more likely to affect protein function than others? Amino acids
- Which ones might be selected against?
- Which ones might be selected for?
- How does this relate to sequence alignment?

Alignments

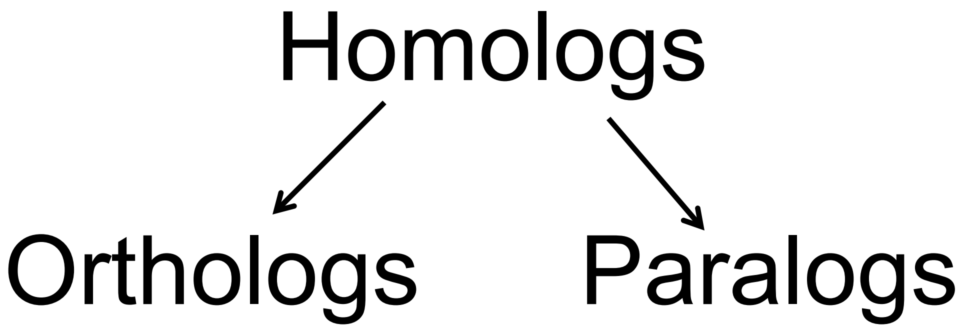
Outline

- Why align sequences?
- Principles of alignments
- Performing alignments
- Scoring alignments: substitution matrices

Why align sequences?

- To determine whether sequences are **homologous**: this is, they are derived from a common ancestor sequence
- Are two sequences so similar that we can conclude they are homologous; or is the similarity just due to chance?
- Databases are so large now that lots of surprising things may occur by chance

Homologs, orthologs, paralogs



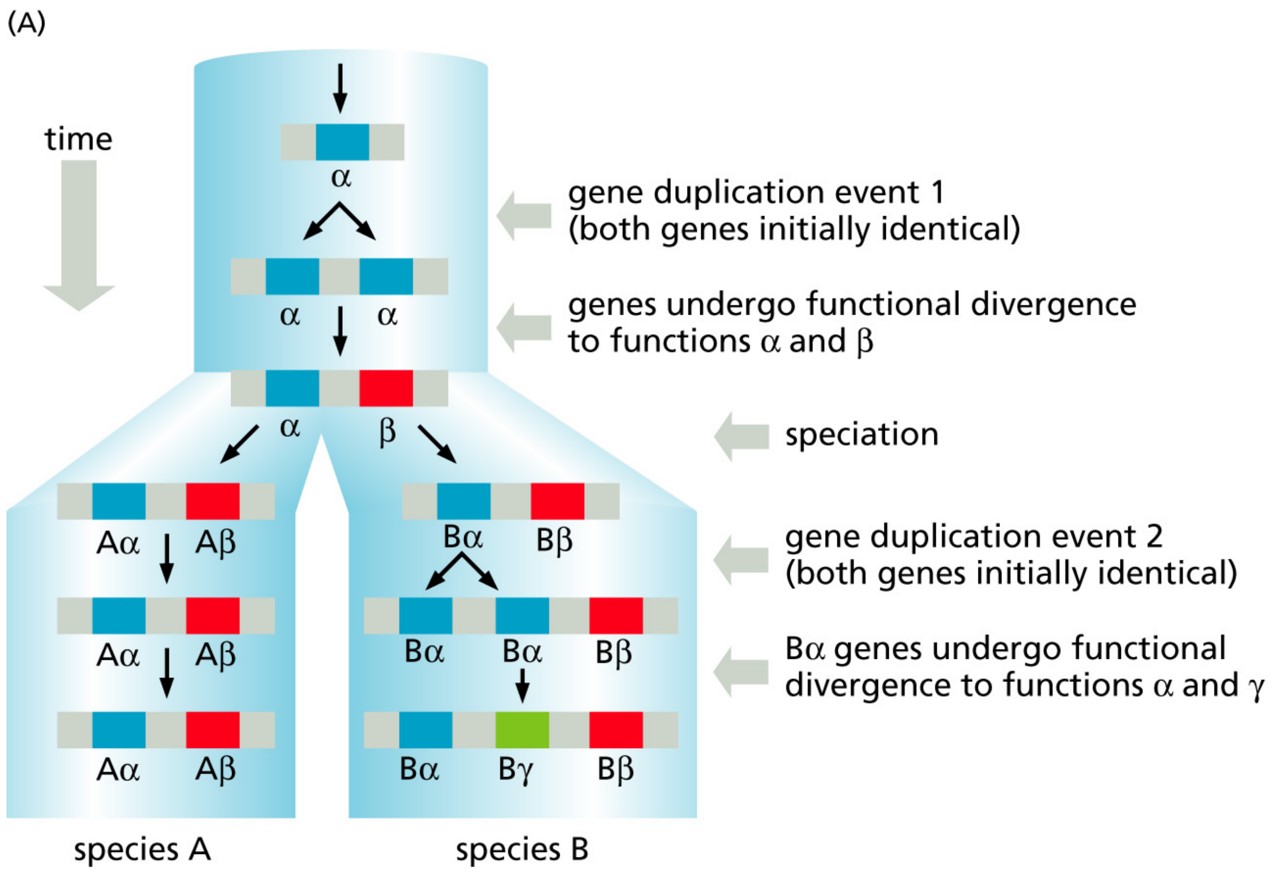
Homologs

Orthologs

Paralogs

- Homologs: Two genes with a common ancester
- Orthologs: Homologous genes arising through **speciation**
- Paralogs: Homologous genes arising through **duplication**

Orthologs vs. paralogs


(A)
time
α
gene duplication event 1
(both genes initially identical)
α α
genes undergo functional divergence
to functions α and β
α β
speciation
Aα Aβ
Bα Bβ
gene duplication event 2
(both genes initially identical)
Aα Aβ
Bα Bα Bβ
Bα genes undergo functional
divergence to functions α and γ
Aα Aβ
Bα Bγ Bβ
species A
species B

All 5 genes are homologs.
Which are orthologs, and which are paralogs?
Which are most likely to function similarly?

Why align sequences?

- To determine whether sequences are **homologous**: this is, they are derived from a common ancestor.
- We hope that **orthologs** will have similar functions.
- Therefore, we make alignments to understand more about how proteins function.
- Alignments also allow us to infer how closely related proteins are
  - Important application: evolution / transmission of disease (e.g. flu, malaria)

Principles of Alignment

- There are almost always multiple ways to align two sequences
- Which way is best? Is the alignment due to more than just chance?
- Need a way to score

Principles of Alignment

```
THISISASEQUENCE
THATSEQUENCE
```

```
THISISASEQUENCE
THAT---SEQUENCE
```

```
THISISASEQUENCE
TH---ATSEQUENCE
```

```
THISISA-SEQUENCE
TH----ATSEQUENCE
```

Which one is best?

Scoring alignments

- Simplest score: % identity
- Number of matches/length of match

```
THISISASEQUENCE
THAT---SEQUENCE
```

10/15 = 66.7%

```
THISISASEQUENCE
TH---ATSEQUENCE
```

10/15 = 66.7%

```
THISISA-SEQUENCE
TH-----ATSEQUENCE
```

11/16 = 68.8%

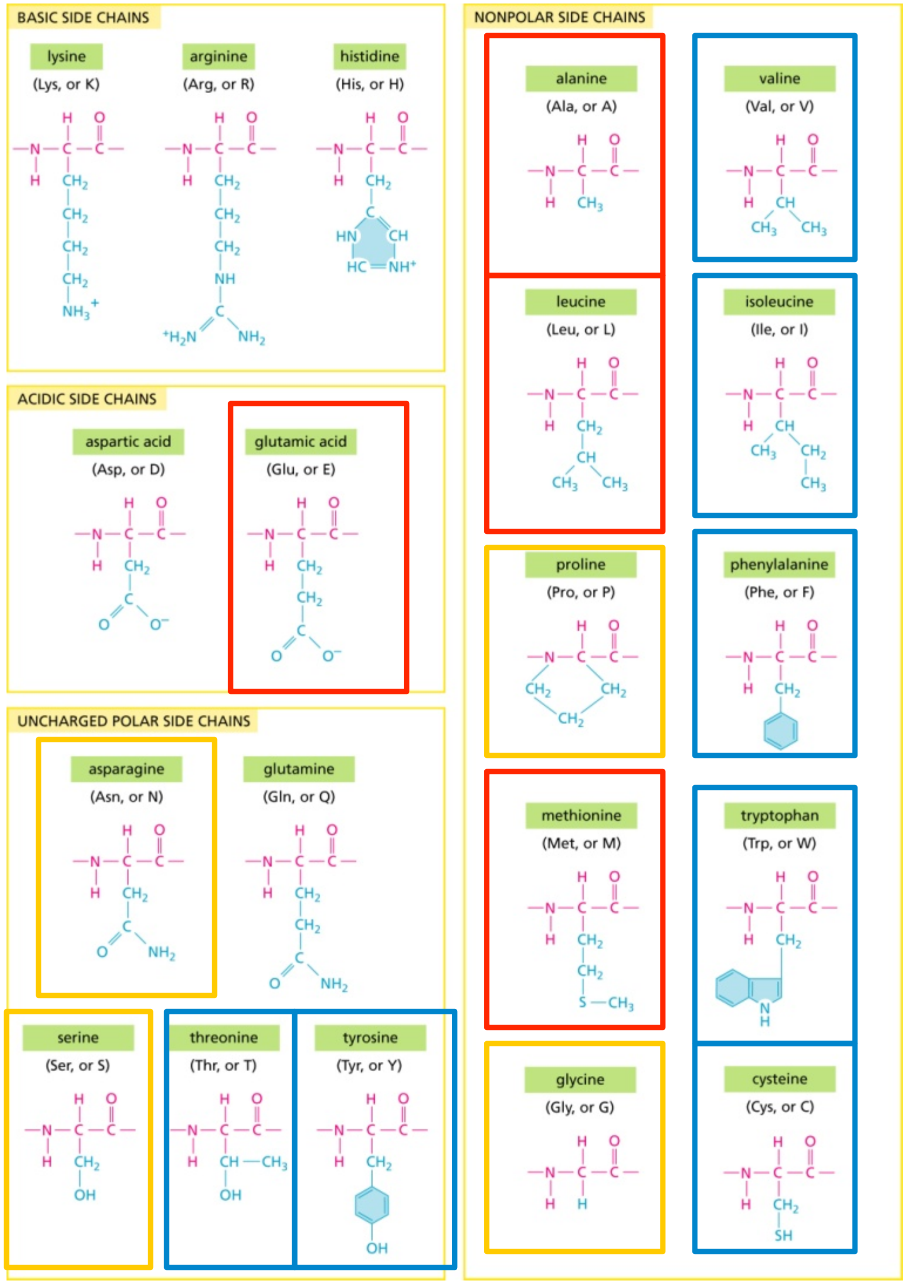
Is the third alignment the best?

Inserting gaps in alignments

- Gaps in alignment presumably due to insertion or deletion in one sequence during evolution
- Too many gaps => meaningless alignment
- Gap penalty
- Gap extension penalty (lower)
- Penalties should be adjusted based on whether you are looking for highly related (high penalty) or more distant sequences (low penalty)
- Gap penalty may depend on residue opposite gap (e.g. tryptophan – large penalty), but typically this is ignored

Identity is not always satisfactory: Substitution matrices

- Use real data to derive scoring matrix
- Genuine matches need not be identical
- How likely is it that an amino acid at a particular position substituted for another amino acid at that position during evolution?
- Substitutions of amino acids with similar physicochemical properties (e.g. size, charge, hydrophobicity) are more likely to conserve function



Amino acids fall into different classes

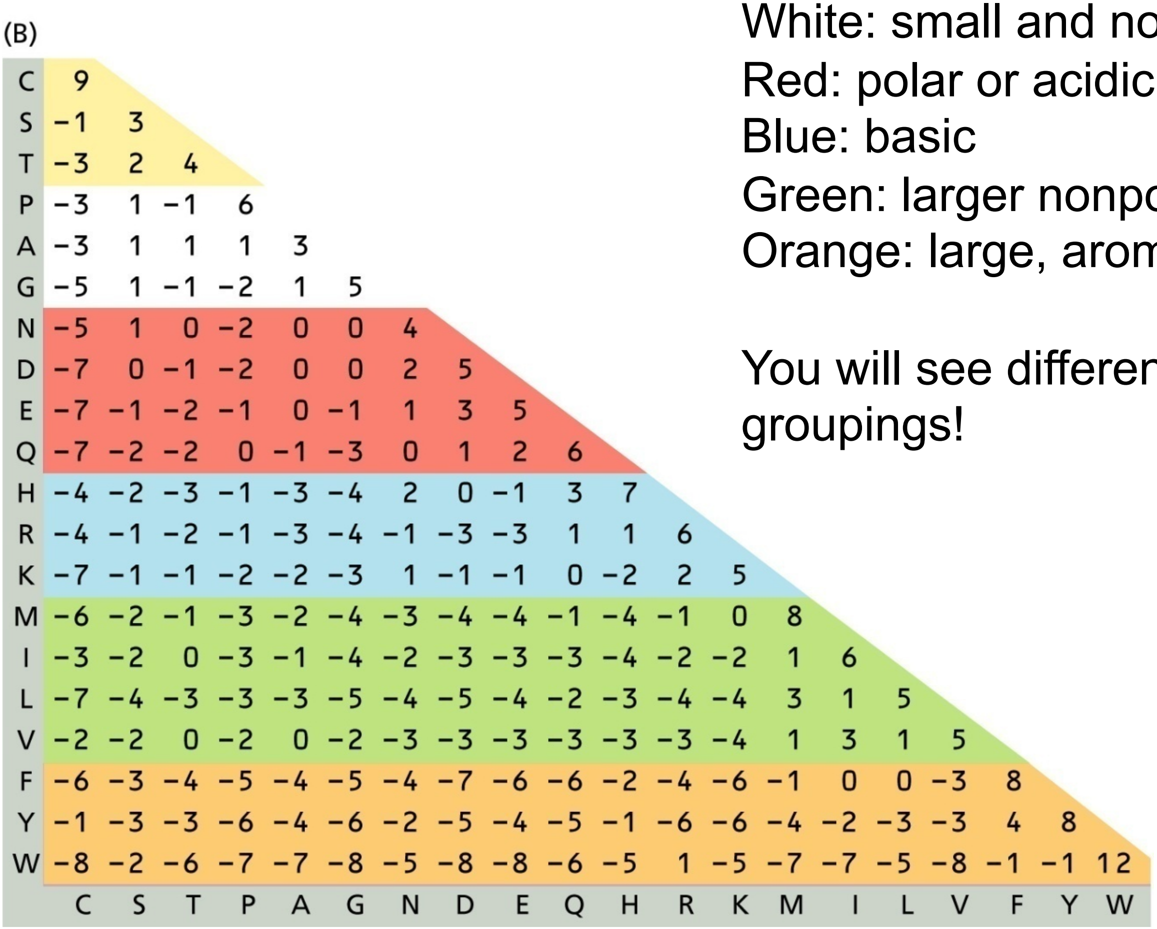
🟥 α-helix formers

🟦 β-strand formers

🟨 Turn segments

Gly, Tyr, and especially Pro are poor α-helix formers

PAM120 Substitution Matrix



(B)

C	9																			
S	−1	3																		
T	−3	2	4																	
P	−3	1	−1	6																
A	−3	1	1	1	3															
G	−5	1	−1	−2	1	5														
N	−5	1	0	−2	0	0	4													
D	−7	0	−1	−2	0	0	2	5												
E	−7	−1	−2	−1	0	−1	1	3	5											
Q	−7	−2	−2	0	−1	−3	0	1	2	6										
H	−4	−2	−3	−1	−3	−4	2	0	−1	3	7									
R	−4	−1	−2	−1	−3	−4	−1	−3	−3	1	1	6								
K	−7	−1	−1	−2	−2	−3	1	−1	−1	0	−2	2	5							
M	−6	−2	−1	−3	−2	−4	−3	−4	−4	−1	−4	−1	0	8						
I	−3	−2	0	−3	−1	−4	−2	−3	−3	−3	−4	−2	−2	1	6					
L	−7	−4	−3	−3	−3	−5	−4	−5	−4	−2	−3	−4	−4	3	1	5				
V	−2	−2	0	−2	0	−2	−3	−3	−3	−3	−3	−3	−4	1	3	1	5			
F	−6	−3	−4	−5	−4	−5	−4	−7	−6	−6	−2	−4	−6	−1	0	0	−3	8		
Y	−1	−3	−3	−6	−4	−6	−2	−5	−4	−5	−1	−6	−6	−4	−2	−3	−3	4	8	
W	−8	−2	−6	−7	−7	−8	−5	−8	−8	−6	−5	1	−5	−7	−7	−5	−8	−1	−1	12
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Yellow: small and polar
White: small and nonpolar/hydrophobic
Red: polar or acidic
Blue: basic
Green: larger nonpolar/hydrophobic
Orange: large, aromatic

You will see different types of amino acid groupings!

Score the alignment:

CSTPEDWLV
CTNCDEWDI

BLAST

- Basic Local Alignment Search Tool
- Most widely used local alignment algorithm

BLAST basics

- Starts with short ‘words’ in the query sequence (default length 3 for proteins, 11 for nucleotides)
- Finds matches in target sequence (using a substitution matrix for proteins, score of match must be above a threshold; for nucleotides, exact match)
- When match is found (two nearby words for proteins), BLAST tries to extend forward and backward to make alignment
- Continues extension until negative scores make the score drop by a critical amount

How do we know an alignment is significant?

- Expect value (E value): ‘the number of times that an alignment as good or better than that found by BLAST would be expected to occur by chance, given the **size of the database searched**’ --From BLAST QuickStart tutorial

$$E = Kmne^{-\lambda S}$$

- S is the score
- Sometimes better to search smaller databases
  - m.n are the lengths of the sequences being compared
  - When comparing to a large database, consider m=query length, n= total length of all sequences in database
- Default E value = 10
- Typically don’t consider matches with E > 0.001
- Often see E values like 10^{-36}

How do we know an alignment is significant?

- Expect value (E value): ‘the number of times that an alignment as good or better than that found by BLAST would be expected to occur by chance, given the **size of the database searched**’ --From BLAST QuickStart tutorial

$$E = Kmne^{-\lambda S}$$

- The parameters *K* and *lambda* can be thought of simply as natural scales for the search space size and the scoring system respectively
- For details: http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html#head2

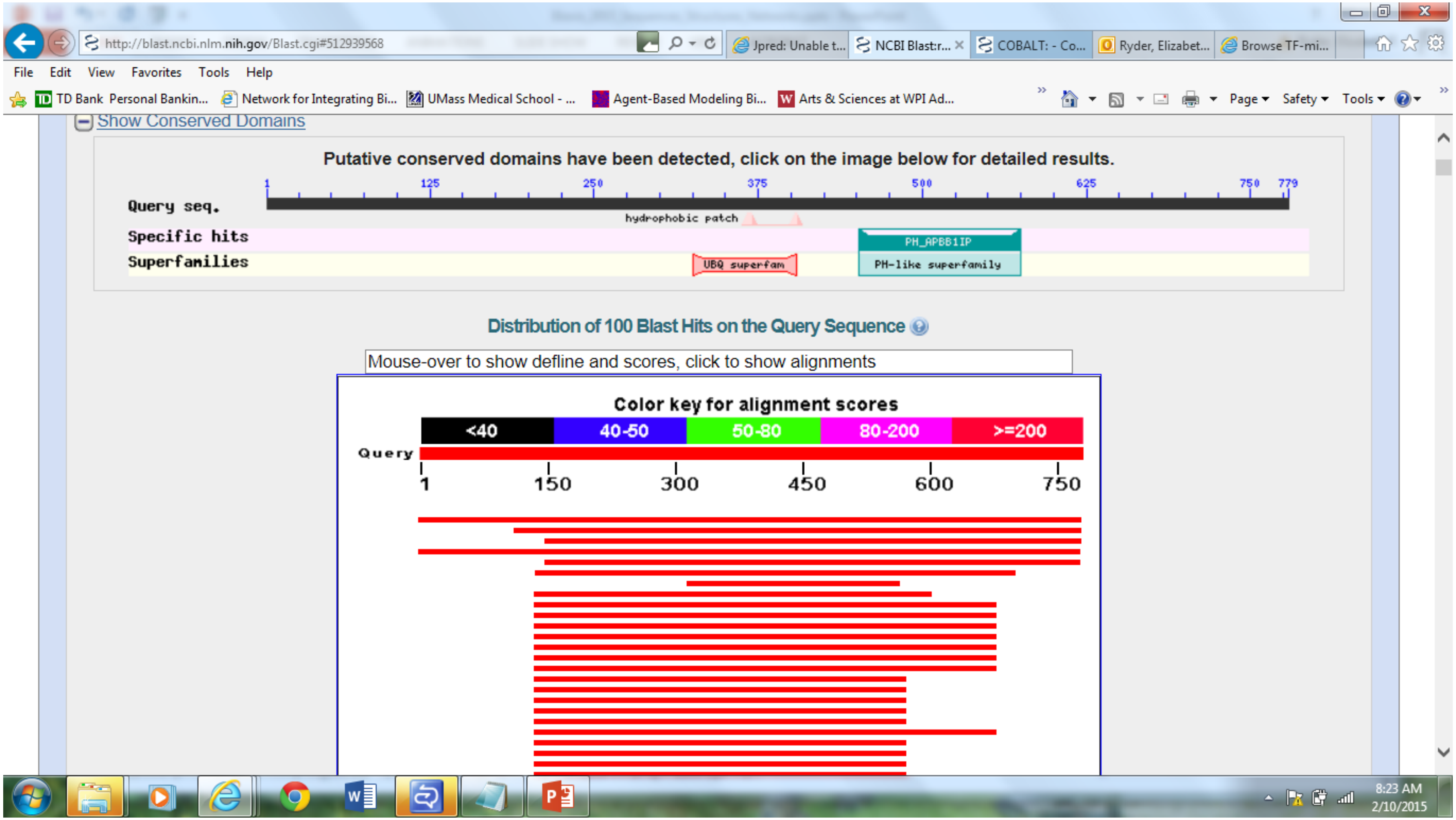
A word about nucleotides

- Why do we typically align proteins and not nucleotide sequences?
- Possible to align nucleotides, but more difficult
  - Only 4 bases
  - Matches more likely to occur by chance
  - Amino acids more conserved over evolution (genetic code is degenerate; DNA less conserved)
- Use protein sequence when available
- Scoring matrices much simpler – e.g. BLASTn uses +2 for a match, -3 for a mismatch
- When would we have to align nucleotides?

Alignment example

- We’ll try aligning the C. elegans protein MIG-10 to the Refseq database
- See if we can decide whether there are homologs of MIG-10 in other species
- Initially, we’ll judge by % identity, % coverage, and E value
- DM0C32V6014 (Expires on 02-11 20:02PM)

BLAST Output



http://blast.ncbi.nlm.nih.gov/Blast.cgi#512939568
Jpred: Unable t...
NCBI Blast:r...
COBALT: - Co...
Ryder, Elizabet...
Browse TF-mi...
File Edit View Favorites Tools Help
TD Bank Personal Bankin...
Network for Integrating Bi...
UMass Medical School - ...
Agent-Based Modeling Bi...
Arts & Sciences at WPI Ad...
Page Safety Tools
8:23 AM
2/10/2015

Show Conserved Domains

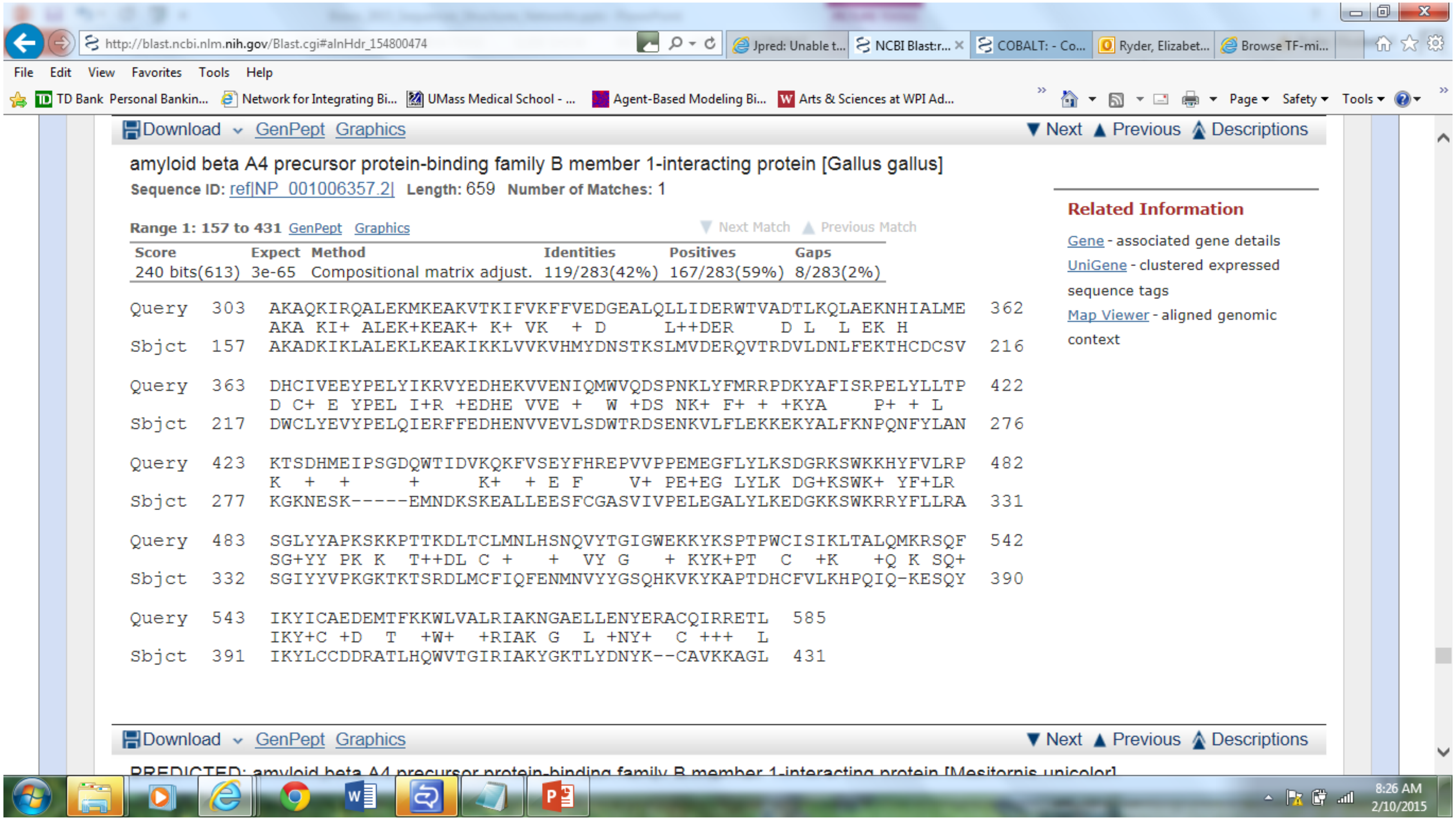
Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Color key for alignment scores

An individual alignment



amyloid beta A4 precursor protein-binding family B member 1-interacting protein [Gallus gallus]

Sequence ID: ref|NP_001006357.2| **Length:** 659 **Number of Matches:** 1

Range 1: 157 to 431 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
240 bits(613)	3e-65	Compositional matrix adjust.	119/283(42%)	167/283(59%)	8/283(2%)

```
Query  303  AKAQKIRQALEKMKEAKVTKIFVKFFVEDGEALQLLIDERWTVADTLKQLAEKNHIALME  362
            AKA KI+ ALEK+KEAK+ K+ VK  + D     L++DER    D L  L EK H
Sbjct  157  AKADKIKLALEKLKEAKIKKLVVKVHMYDNSTKSLMVDERQVTRDVLDNLFEKTHCDCSV  216

Query  363  DHCIVEEYPELYIKRVYEDHEKVVENIQMWVQDSPNKLYFMRRPDKYAFISRPELYLLTP  422
            D C+ E YPEL I+R +EDHE VVE +  W +DS NK+ F+ + +KYA    P+ + L
Sbjct  217  DWCLYEVYPELQIERFFEDHENVVEVLSDWTRDSENKVLFLEKKEKYALFKNPQNFYLAN  276

Query  423  KTSDHMEIPSGDQWTIDVKQKFVSEYFHREPVVPPEMEGFLYLKSDGRKSWKKHYFVLRP  482
            K  +  +     +     K+  + E F    V+ PE+EG LYLK DG+KSWK+ YF+LR
Sbjct  277  KGKNESK-----EMNDKSKEALLEESFCGASVIVPELEGALYLKEDGKKSWKRRYFLLRA  331

Query  483  SGLYYAPKSKKPTTKDLTCLMNLHSNQVYTGIGWEKKYKSPTPWCISIKLTALQMKRSQF  542
            SG+YY PK K  T++DL C +   +  VY G   + KYK+PT  C  +K   +Q K SQ+
Sbjct  332  SGIYYVPKGKTKTSRDLMCFIQFENMNVYYGSQHKVKYKAPTDHCFVLKHPQIQ-KESQY  390

Query  543  IKYICAEDEMTFKKWLVALRIAKNGAELLENYERACQIRRETL  585
            IKY+C +D  T  +W+  +RIAK G  L +NY+  C +++  L
Sbjct  391  IKYLCCDDRATLHQWVTGIRIAKYGKTLYDNYK--CAVKKAGL  431
```

Related Information

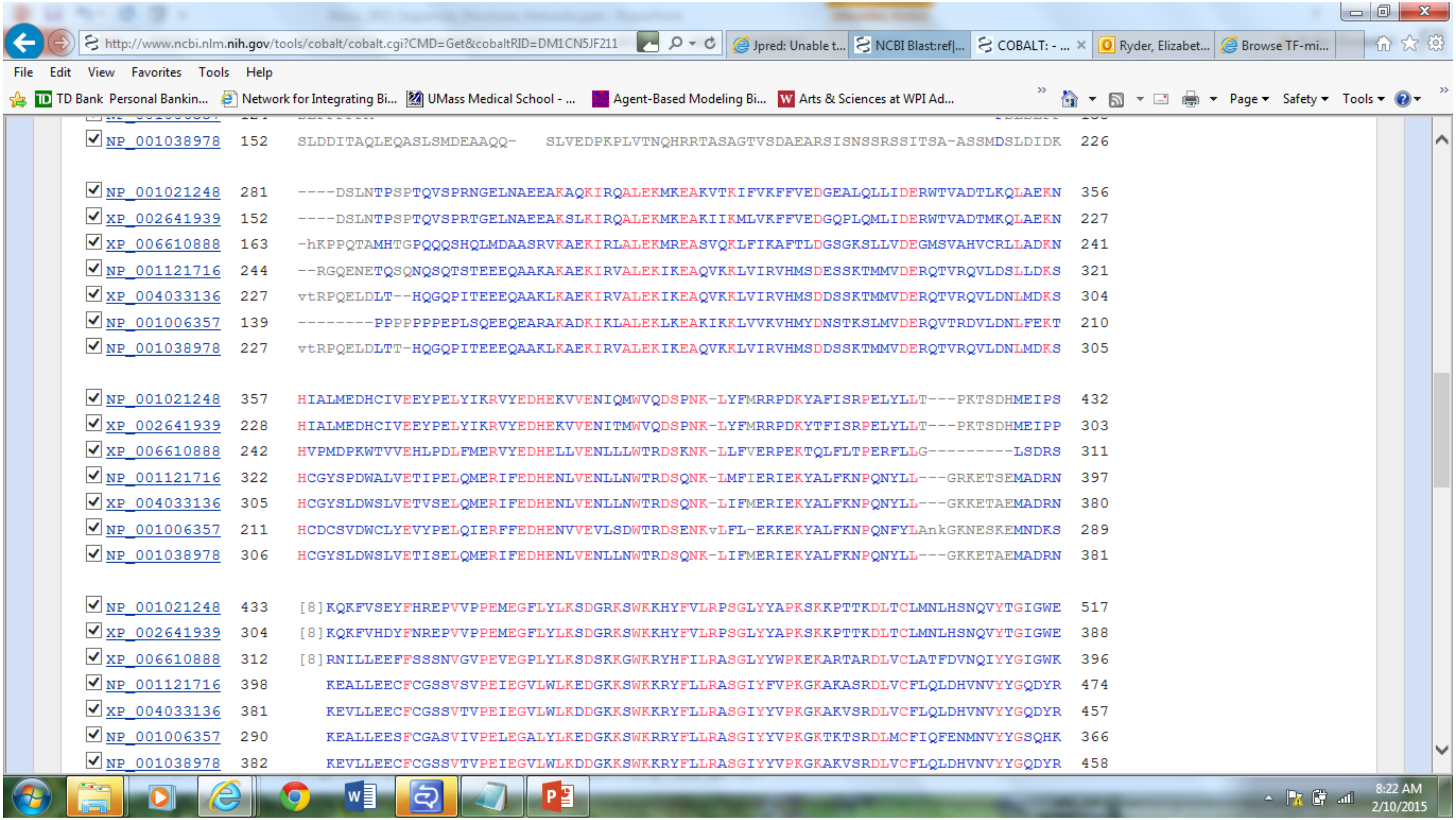
Gene - associated gene details

UniGene - clustered expressed sequence tags

Map Viewer - aligned genomic context

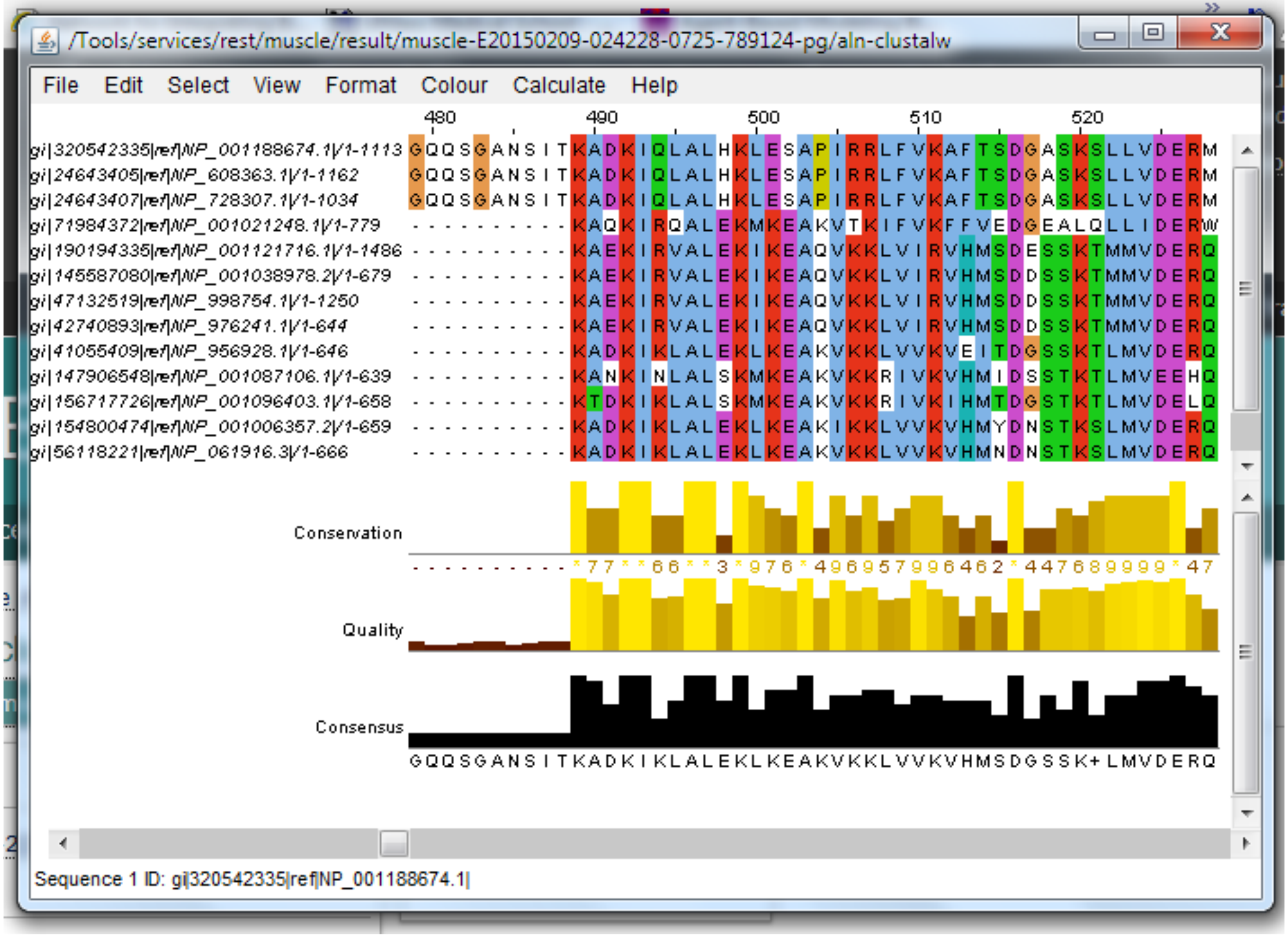
PREDICTED: amyloid beta A4 precursor protein-binding family B member 1-interacting protein [Mesitornis unicolor]

Multiple alignment: COBALT Output (NCBI)


http://www.ncbi.nlm.nih.gov/tools/cobalt/cobalt.cgi?CMD=Get&cobaltRID=DM1CN5JF211
Jpred: Unable t...
NCBI Blast:ref|...
COBALT: - ...
Ryder, Elizabet...
Browse TF-mi...
File Edit View Favorites Tools Help
TD Bank Personal Bankin...
Network for Integrating Bi...
UMass Medical School - ...
Agent-Based Modeling Bi...
Arts & Sciences at WPI Ad...
Page Safety Tools
8:22 AM
2/10/2015

```
☑ NP_001038978  152  SLDDITAQLEQASLSMDEAAQQ-    SLVEDPKPLVTNQHRRTASAGTVSDAEARSISNSSRSSITSA-ASSMDSLDIDK  226

☑ NP_001021248  281  ----DSLNTPSPTQVSPRNGELNAEEAKAQKIRQALEKMKEAKVTKIFVKFFVEDGEALQLLIDERWTVADTLKQLAEKN  356
☑ XP_002641939  152  ----DSLNTPSPTQVSPRTGELNAEEAKSLKIRQALEKMKEAKIIKMLVKFFVEDGQPLQMLIDERWTVADTMKQLAEKN  227
☑ XP_006610888  163  -hKPPQTAMHTGPQQQSHQLMDAASRVKAEKIRLALEKMREASVQKLFIKAFTLDGSGKSLLVDEGMSVAHVCRLLADKN  241
☑ NP_001121716  244  --RGQENETQSQNQSQTSTEEEQAAKAKAEKIRVALEKIKEAQVKKLVIRVHMSDESSKTMMVDERQTVRQVLDSLLDKS  321
☑ XP_004033136  227  vtRPQELDLT--HQGQPITEEEQAAKLKAEKIRVALEKIKEAQVKKLVIRVHMSDDSSKTMMVDERQTVRQVLDNLMDKS  304
☑ NP_001006357  139  --------PPPPPPPEPLSQEEQEARAKADKIKLALEKLKEAKIKKLVVKVHMYDNSTKSLMVDERQVTRDVLDNLFEKT  210
☑ NP_001038978  227  vtRPQELDLTT-HQGQPITEEEQAAKLKAEKIRVALEKIKEAQVKKLVIRVHMSDDSSKTMMVDERQTVRQVLDNLMDKS  305

☑ NP_001021248  357  HIALMEDHCIVEEYPELYIKRVYEDHEKVVENIQMWVQDSPNK-LYFMRRPDKYAFISRPELYLLT---PKTSDHMEIPS  432
☑ XP_002641939  228  HIALMEDHCIVEEYPELYIKRVYEDHEKVVENITMWVQDSPNK-LYFMRRPDKYTFISRPELYLLT---PKTSDHMEIPP  303
☑ XP_006610888  242  HVPMDPKWTVVEHLPDLFMERVYEDHELLVENLLLWTRDSKNK-LLFVERPEKTQLFLTPERFLLG---------LSDRS  311
☑ NP_001121716  322  HCGYSPDWALVETIPELQMERIFEDHENLVENLLNWTRDSQNK-LMFIERIEKYALFKNPQNYLL---GRKETSEMADRN  397
☑ XP_004033136  305  HCGYSLDWSLVETVSELQMERIFEDHENLVENLLNWTRDSQNK-LIFMERIEKYALFKNPQNYLL---GKKETAEMADRN  380
☑ NP_001006357  211  HCDCSVDWCLYEVYPELQIERFFEDHENVVEVLSDWTRDSENKvLFL-EKKEKYALFKNPQNFYLAnkGKNESKEMNDKS  289
☑ NP_001038978  306  HCGYSLDWSLVETISELQMERIFEDHENLVENLLNWTRDSQNK-LIFMERIEKYALFKNPQNYLL---GKKETAEMADRN  381

☑ NP_001021248  433  [8]KQKFVSEYFHREPVVPPEMEGFLYLKSDGRKSWKKHYFVLRPSGLYYAPKSKKPTTKDLTCLMNLHSNQVYTGIGWE  517
☑ XP_002641939  304  [8]KQKFVHDYFNREPVVPPEMEGFLYLKSDGRKSWKKHYFVLRPSGLYYAPKSKKPTTKDLTCLMNLHSNQVYTGIGWE  388
☑ XP_006610888  312  [8]RNILLEEFFSSSNVGVPEVEGPLYLKSDSKKGWKRYHFILRASGLYYWPKEKARTARDLVCLATFDVNQIYYGIGWK  396
☑ NP_001121716  398     KEALLEECFCGSSVSVPEIEGVLWLKEDGKKSWKKRYFLLRASGIYFVPKGKAKASRDLVCFLQLDHVNVYYGQDYR  474
☑ XP_004033136  381     KEVLLEECFCGSSVTVPEIEGVLWLKDDGKKSWKKRYFLLRASGIYYVPKGKAKVSRDLVCFLQLDHVNVYYGQDYR  457
☑ NP_001006357  290     KEALLEESFCGASVIVPELEGALYLKEDGKKSWKRRYFLLRASGIYYVPKGKTKTSRDLMCFIQFENMNVYYGSQHK  366
☑ NP_001038978  382     KEVLLEECFCGSSVTVPEIEGVLWLKDDGKKSWKKRYFLLRASGIYYVPKGKAKVSRDLVCFLQLDHVNVYYGQDYR  458
```

MIG-10 in Jalview (from EBI)

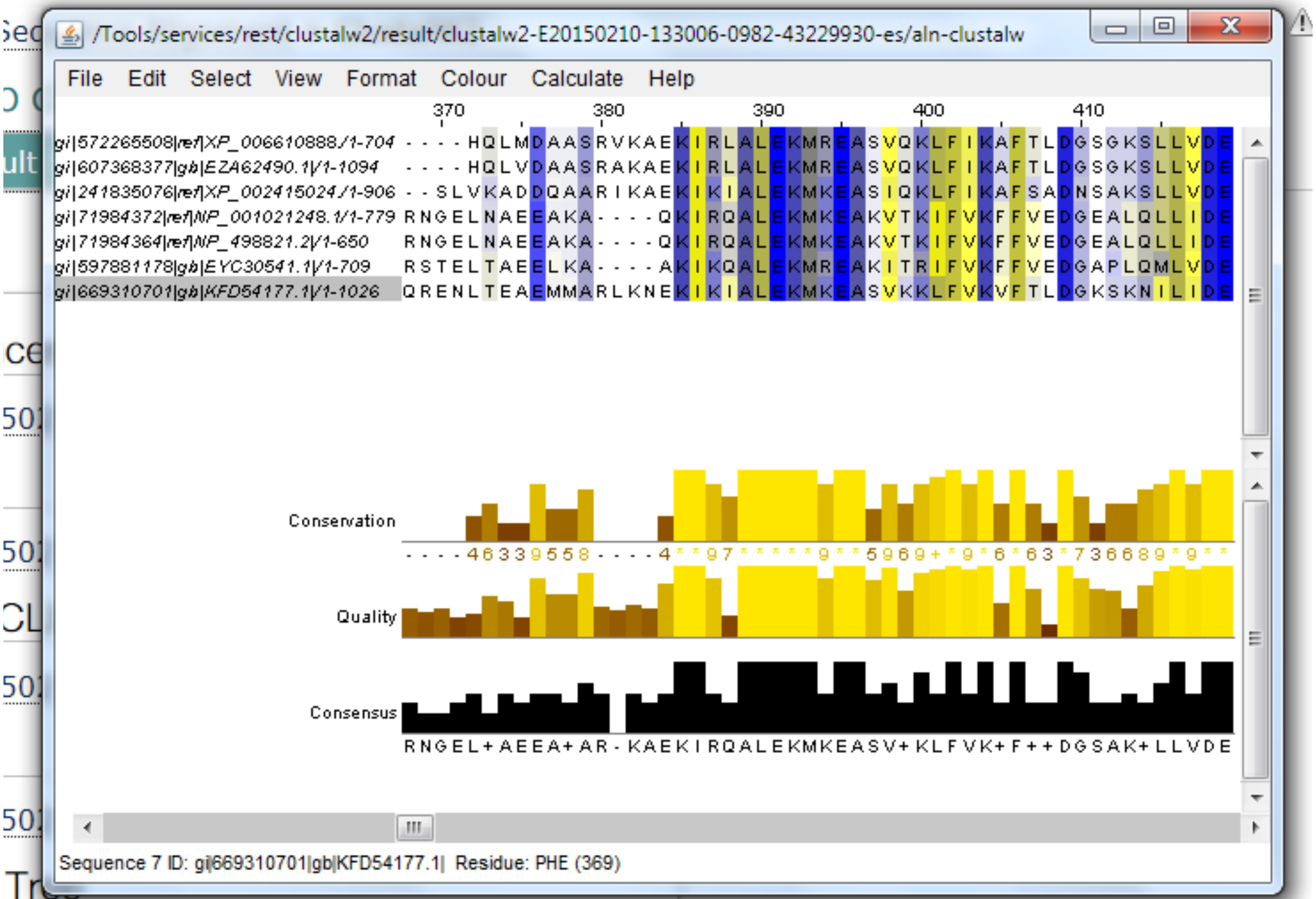

/Tools/services/rest/muscle/result/muscle-E20150209-024228-0725-789124-pg/aln-clustalw
File
Edit
Select
View
Format
Colour
Calculate
Help

```
                                          480       490       500       510       520
gi|320542335|ref|NP_001188674.1|/1-1113  GQQSGANSITKADKIQLALHKLESAPIRRLFVKAFTSDGASKSLLVDERM
gi|24643405|ref|NP_608363.1|/1-1162      GQQSGANSITKADKIQLALHKLESAPIRRLFVKAFTSDGASKSLLVDERM
gi|24643407|ref|NP_728307.1|/1-1034      GQQSGANSITKADKIQLALHKLESAPIRRLFVKAFTSDGASKSLLVDERM
gi|71984372|ref|NP_001021248.1|/1-779    ----------KAQKIRQALEKMKEAKVTKIFVKFFVEDGEALQLLIDERW
gi|190194335|ref|NP_001121716.1|/1-1486  ----------KAEKIRVALEKIKEAQVKKLVIRVHMSDESSKTMMVDERQ
gi|145587080|ref|NP_001038978.2|/1-679   ----------KAEKIRVALEKIKEAQVKKLVIRVHMSDDSSKTMMVDERQ
gi|47132519|ref|NP_998754.1|/1-1250      ----------KAEKIRVALEKIKEAQVKKLVIRVHMSDDSSKTMMVDERQ
gi|42740893|ref|NP_976241.1|/1-644       ----------KAEKIRVALEKIKEAQVKKLVIRVHMSDDSSKTMMVDERQ
gi|41055409|ref|NP_956928.1|/1-646       ----------KADKIKLALEKLKEAKVKKLVVKVEITDGSSKTLMVDERQ
gi|147906548|ref|NP_001087106.1|/1-639   ----------KANKINLALSKMKEAKVKKRIVKVHMIDSSTKTLMVEEHQ
gi|156717726|ref|NP_001096403.1|/1-658   ----------KTDKIKLALSKMKEAKVKKRIVKIHMTDGSTKTLMVDELQ
gi|154800474|ref|NP_001006357.2|/1-659   ----------KADKIKLALEKLKEAKIKKLVVKVHMYDNSTKSLMVDERQ
gi|56118221|ref|NP_061916.3|/1-666       ----------KADKIKLALEKLKEAKVKKLVVKVHMNDNSTKSLMVDERQ

                           Conservation
                                         ----------*77**66**3*976*496957996462*447689999*47

                                Quality

                              Consensus
                                         GQQSGANSITKADKIKLALEKLKEAKVKKLVVKVHMSDGSSK+LMVDERQ
```

Sequence 1 ID: gi|320542335|ref|NP_001188674.1|

http://www.ebi.ac.uk/Tools/services/web_clustalw2/toolresult.ebi?jobId=clustalw2-E20150210-133006-0982-43229930-es

Jalview colored by conservation



/Tools/services/rest/clustalw2/result/clustalw2-E20150210-133006-0982-43229930-es/aln-clustalw
File
Edit
Select
View
Format
Colour
Calculate
Help
370
380
390
400
410
Conservation
Quality
Consensus

```
gi|572265508|ref|XP_006610888/1-704   ----HQLMDAASRVKAEKIRLALEKMREASVQKLFIKAFTLDGSGKSLLVDE
gi|607368377|gb|EZA62490.1|/1-1094    ----HQLVDAASRAKAEKIRLALEKMREASVQKLFIKAFTLDGSGKSLLVDE
gi|241835076|ref|XP_002415024/1-906   --SLVKADDQAARIKAEKIKIALEKMKEASIQKLFIKAFSADNSAKSLLVDE
gi|71984372|ref|NP_001021248.1/1-779  RNGELNAEEAKA----QKIRQALEKMKEAKVTKIFVKFFVEDGEALQLLIDE
gi|71984364|ref|NP_498821.2|/1-650    RNGELNAEEAKA----QKIRQALEKMKEAKVTKIFVKFFVEDGEALQLLIDE
gi|597881178|gb|EYC30541.1|/1-709     RSTELTAEELKA----AKIKQALEKMREAKITRIFVKFFVEDGAPLQMLVDE
gi|669310701|gb|KFD54177.1|/1-1026    QRENLTEAEMMARLKNEKIKIALEKMKEASVKKLFVKVFTLDGKSKNILIDE

Conservation                          ----46339558----4**97*****9**5969+*9*6*63*736689*9**
Quality
Consensus                             RNGEL+AEEA+AR-KAEKIRQALEKMKEASV+KLFVK+F++DGSAK+LLVDE
```

Sequence 7 ID: gi|669310701|gb|KFD54177.1| Residue: PHE (369)

Gene regulatory networks

How is gene expression controlled?

Prokaryotic and eukaryotic transcription and translation

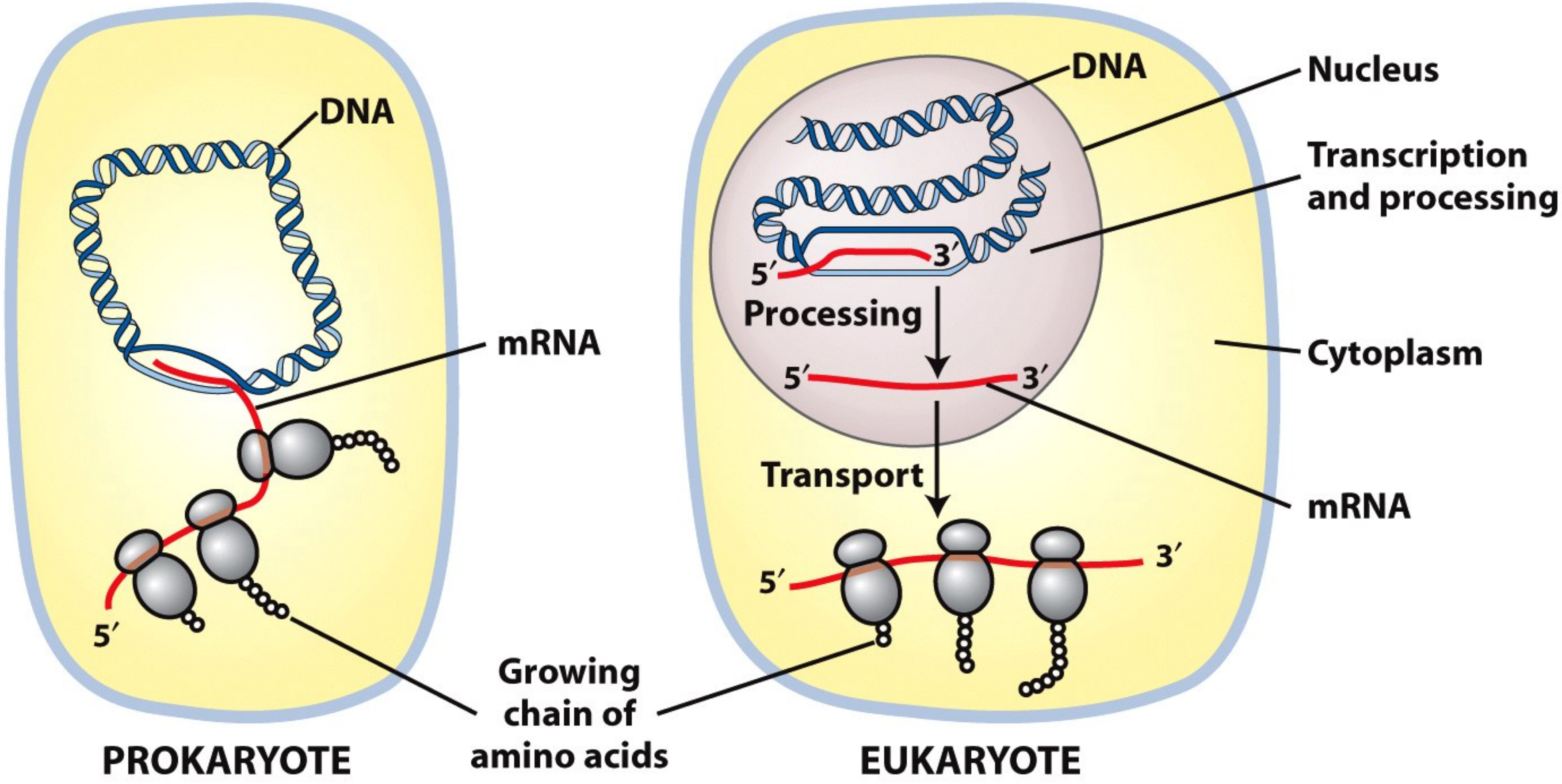
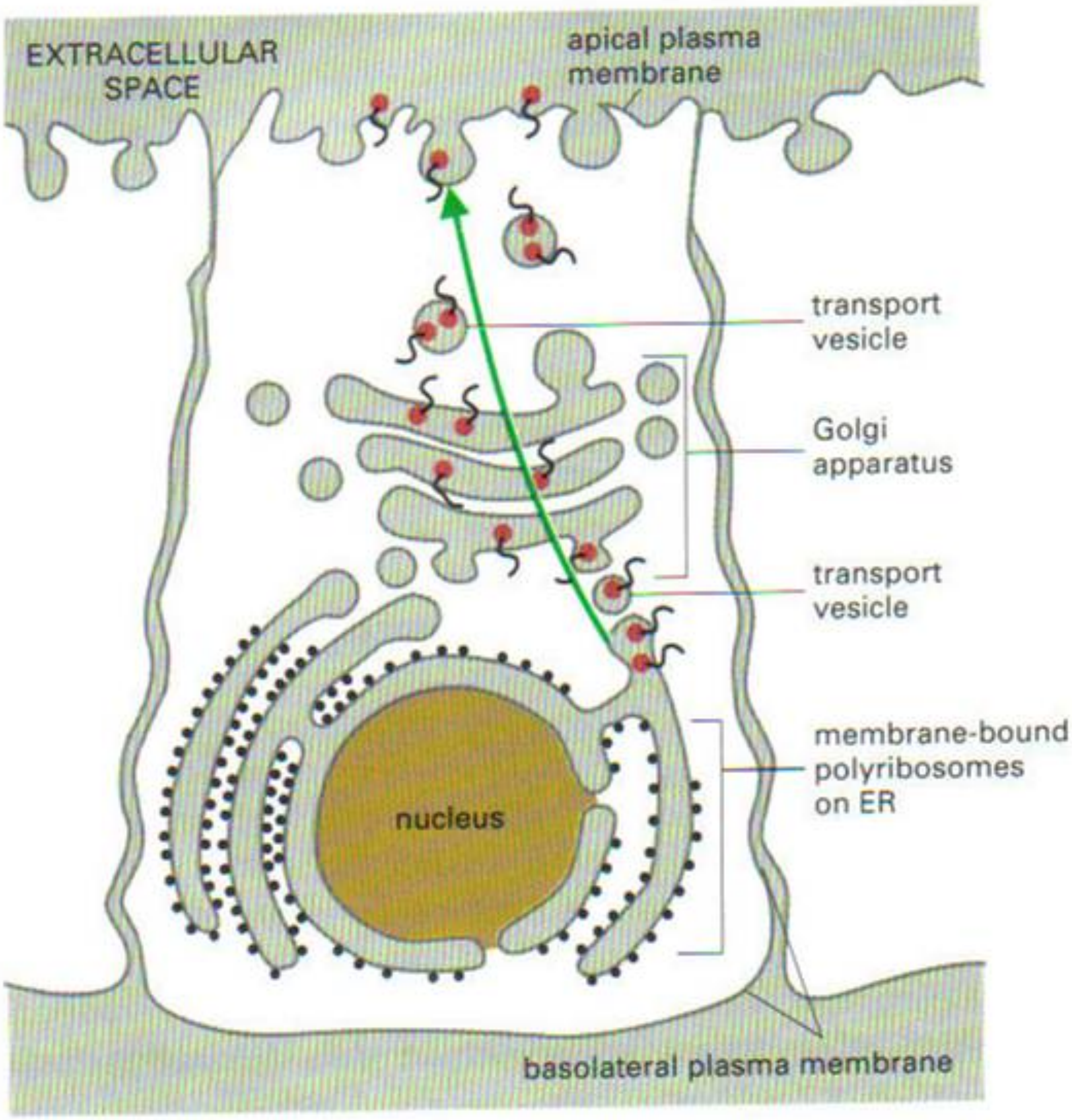


Figure 8-11
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

Griffeths, Introduction to Genetic Analysis, 2008

Topology of glycosylation



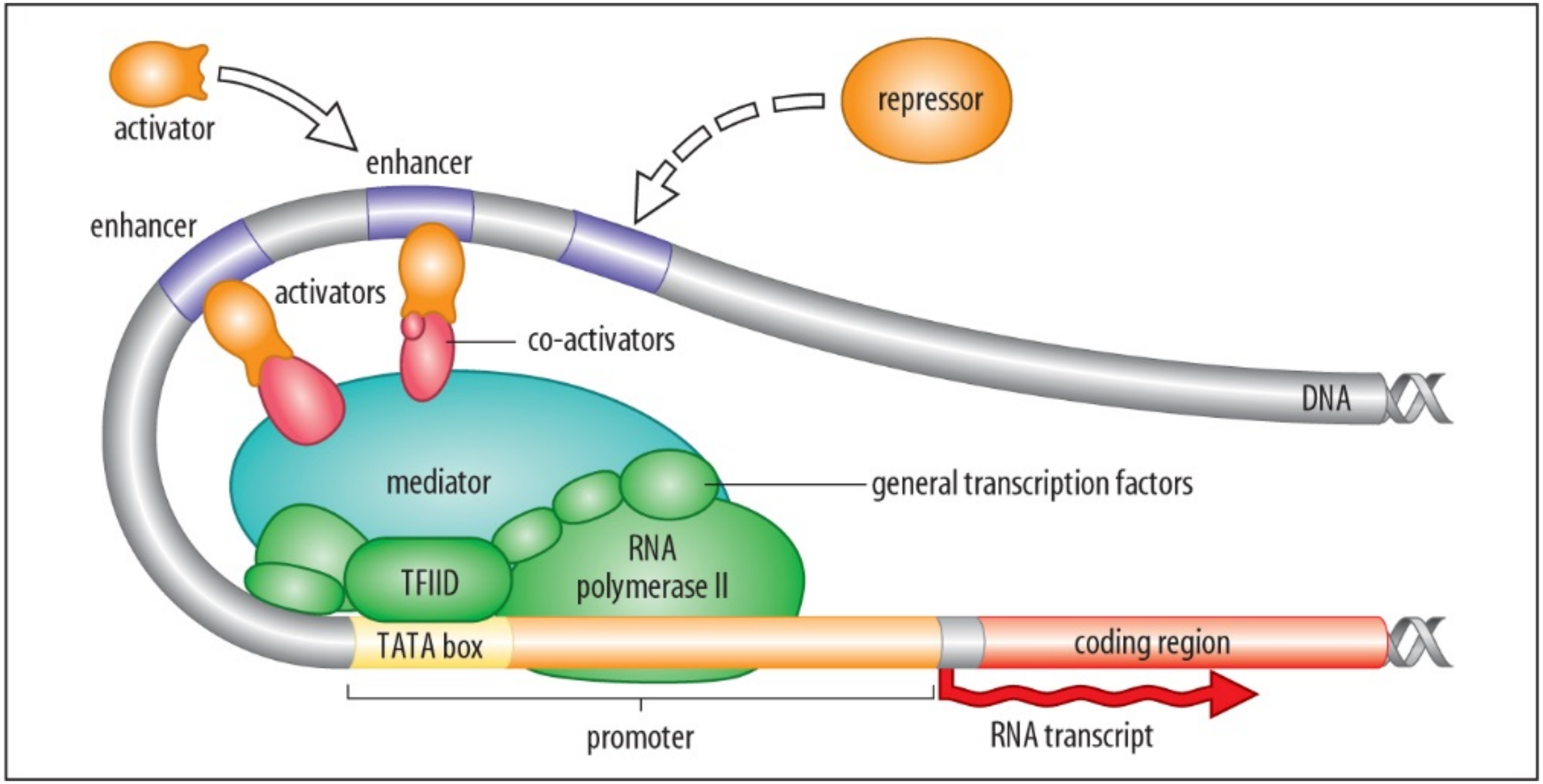
∫ membrane protein

● N-linked oligosaccharide added in ER

Glycosylation and phosphorylation occur in different cellular compartments…

Alberts et al., Molecular Biology of the Cell

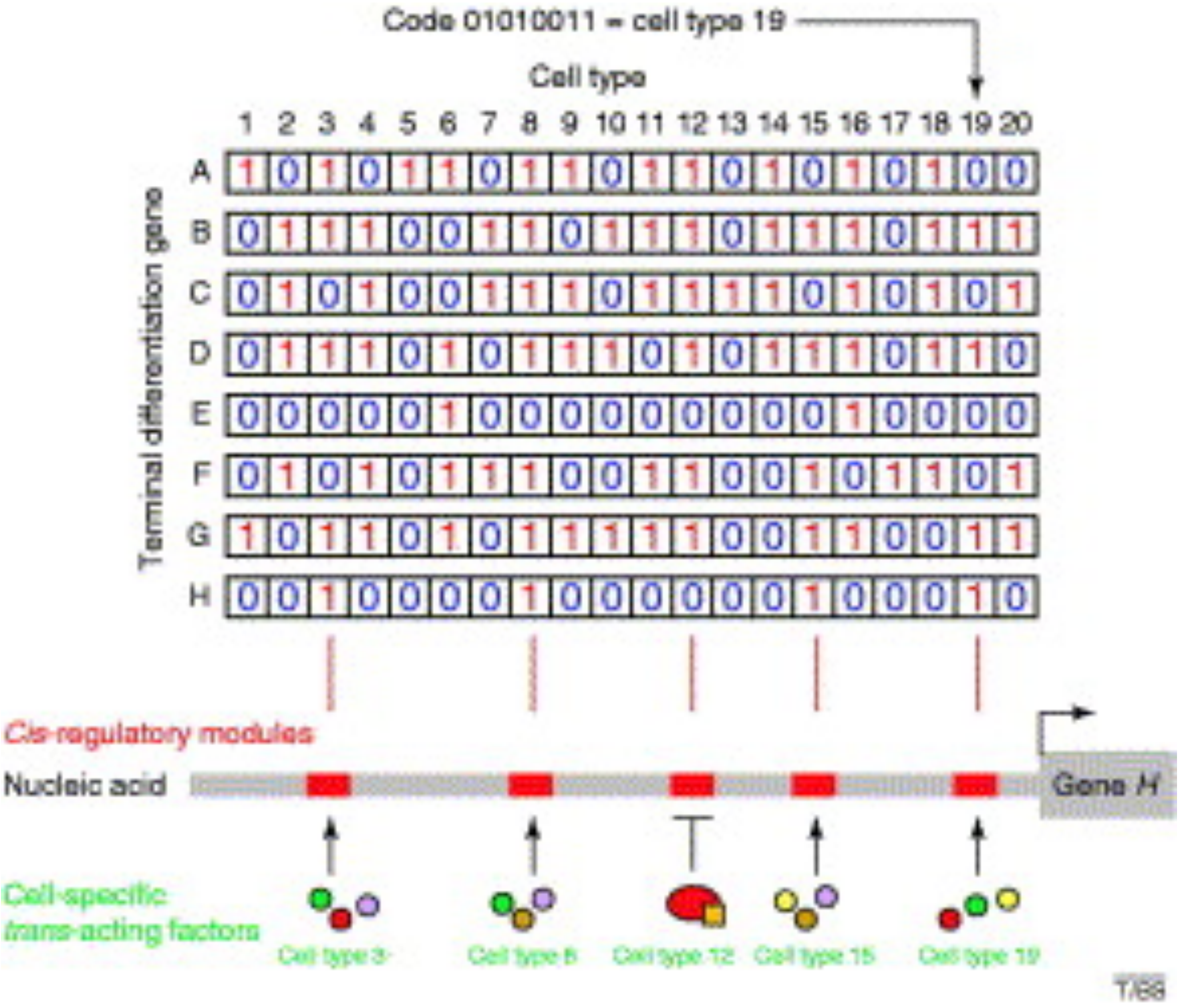
Eukaryotic transcription is complex!



- Basal transcriptional regulators
- Cell type specific enhancers and repressors

http://www.mun.ca/biology/desmid/brian/BIOL3530/DEVO_10/devo_10.html

‘Gene batteries’



Code 01010011 = cell type 19

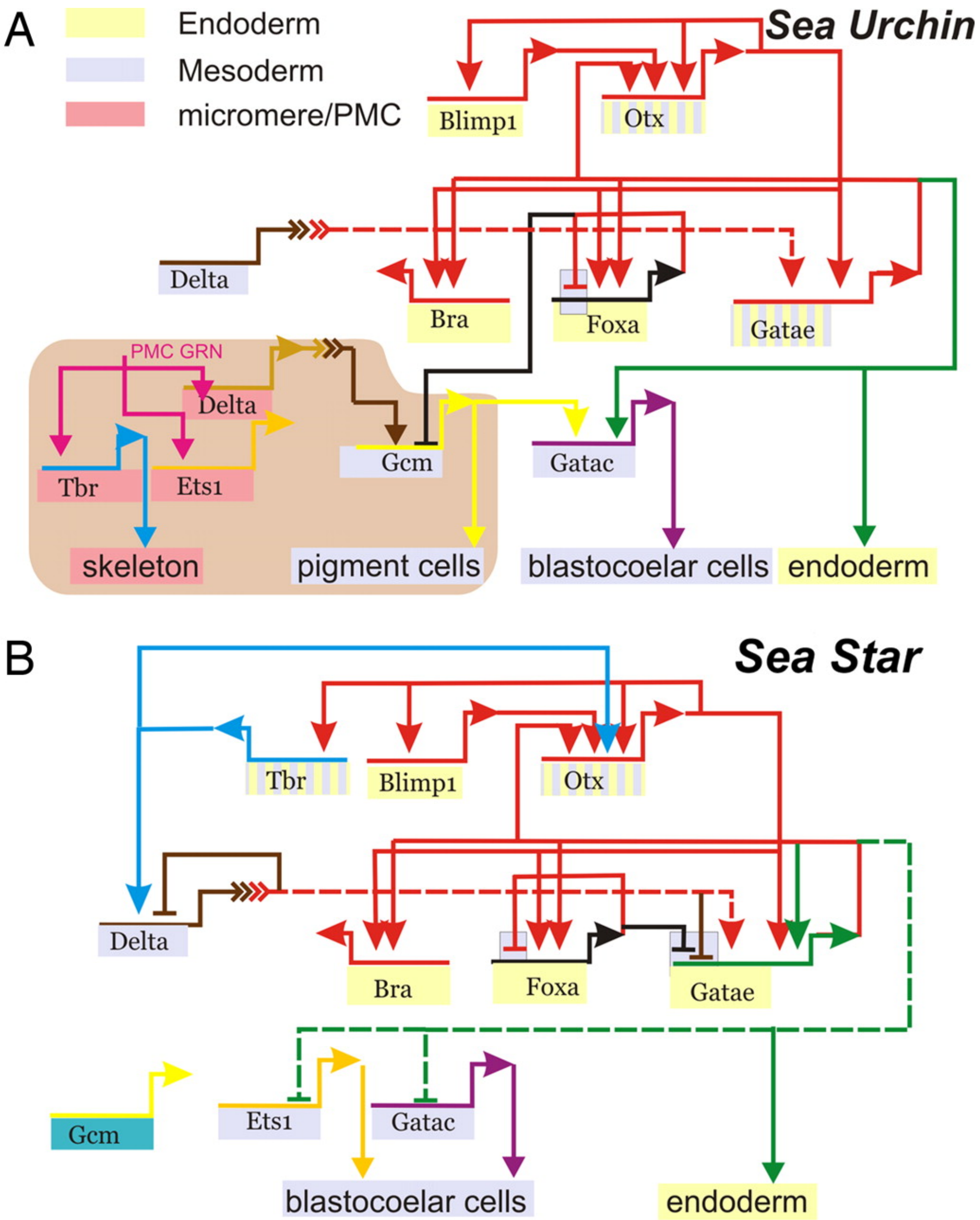
Cell type

Terminal differentiation gene	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	1	0	1	0	1	1	0	1	1	0	1	1	0	1	0	1	0	1	0	0
B	0	1	1	1	0	0	1	1	0	1	1	1	0	1	1	1	0	1	1	1
C	0	1	0	1	0	0	1	1	1	0	1	1	1	1	0	1	0	1	0	1
D	0	1	1	1	0	1	0	1	1	1	0	1	0	1	1	1	0	1	1	0
E	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
F	0	1	0	1	0	1	1	1	0	0	1	1	0	0	1	0	1	1	0	1
G	1	0	1	1	0	1	0	1	1	1	1	1	0	0	1	1	0	0	1	1
H	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0

Terminal differentiation genes expressed by different cell types

Hobert. Trends in Biochemical Sciences, Volume 29, Issue 9, 2004, 462 – 468. http://dx.doi.org/10.1016/j.tibs.2004.07.001

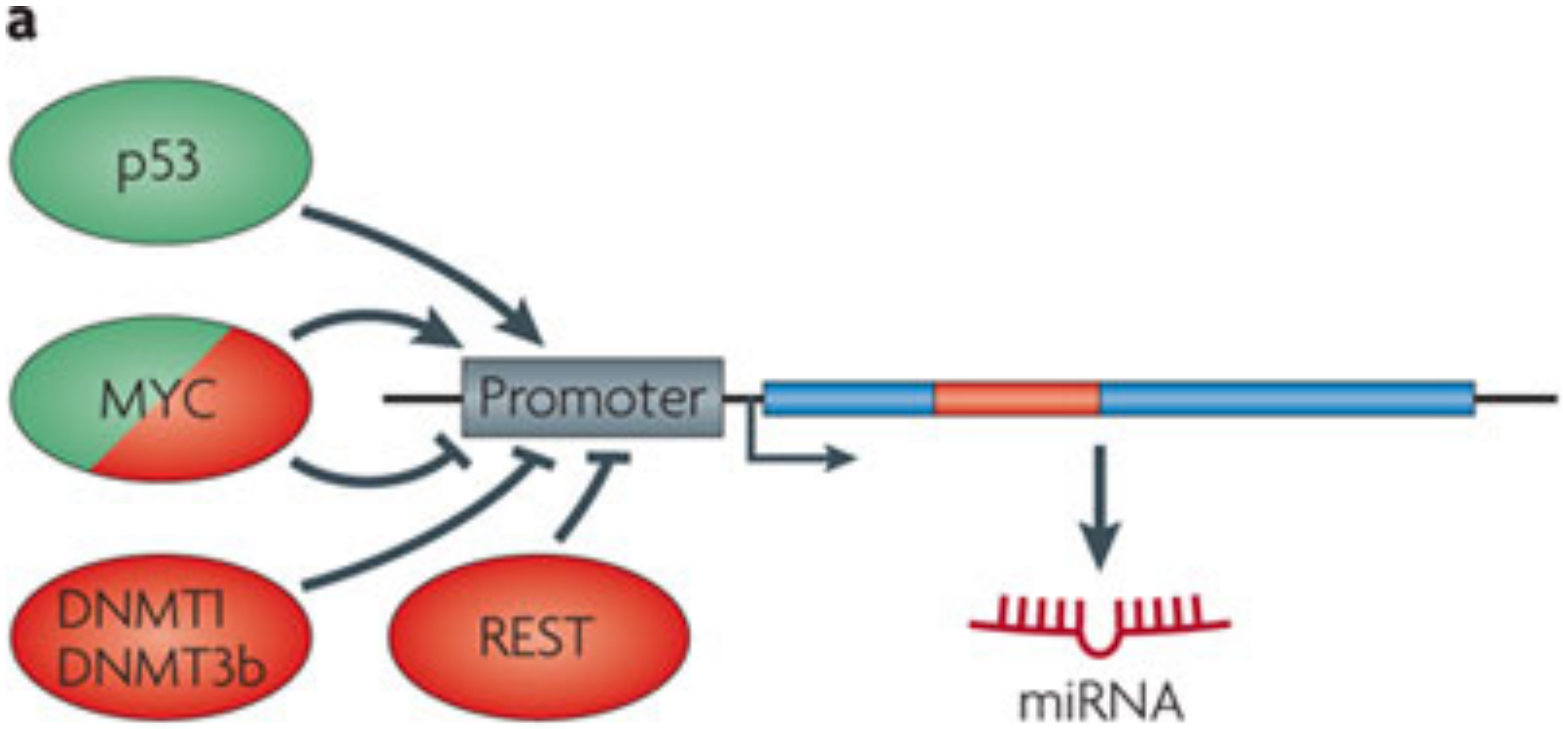
GRN: Sea urchin vs. Sea star



- All genes except Delta are transcription factors
- Arrows: + regulation
- T's: - regulation
- Colors: ???

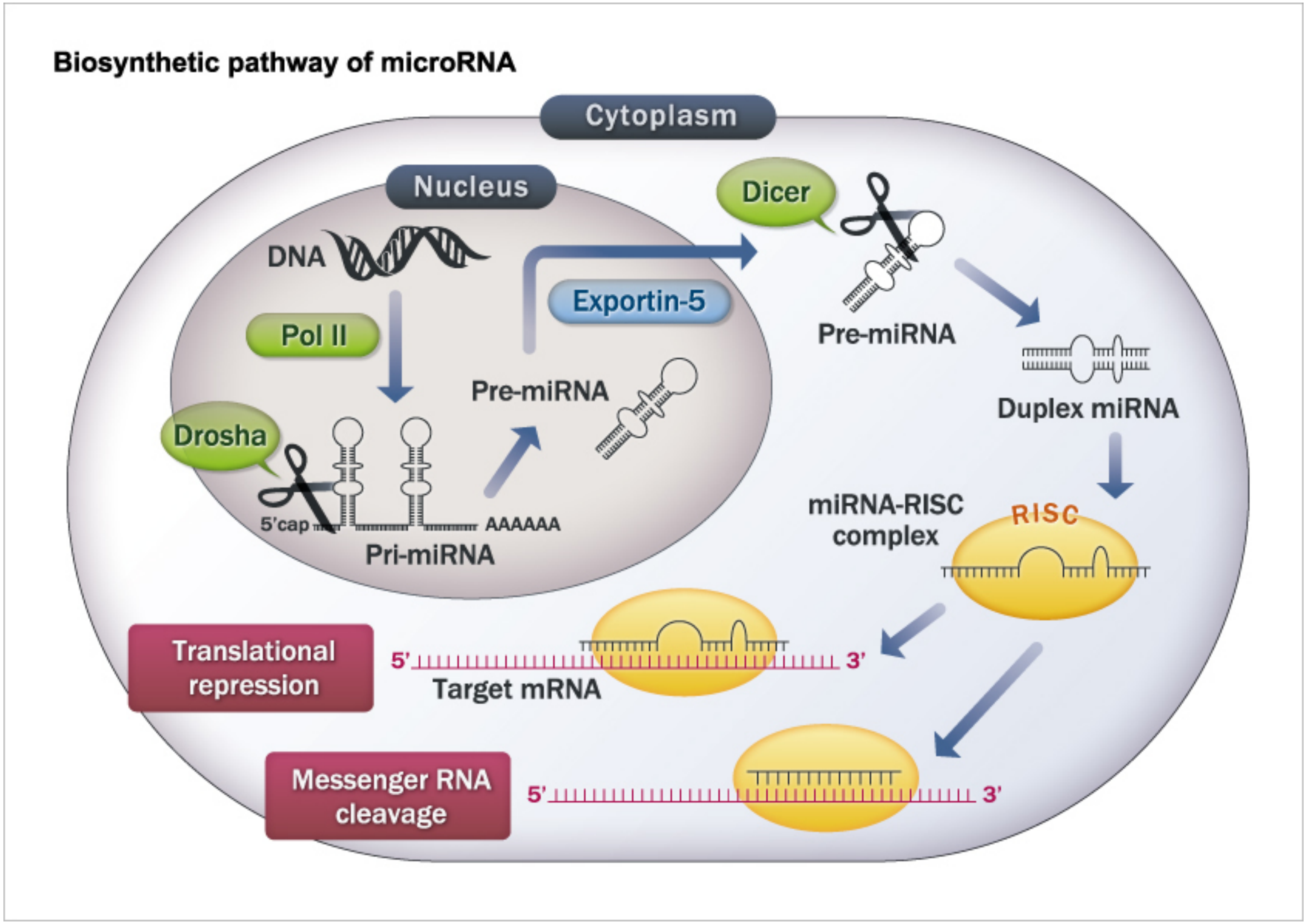
http://www.pnas.org/content/104/49/19404/F5.expansion.html

Micro RNA (miRNA) genes are regulated similarly to protein-encoding genes

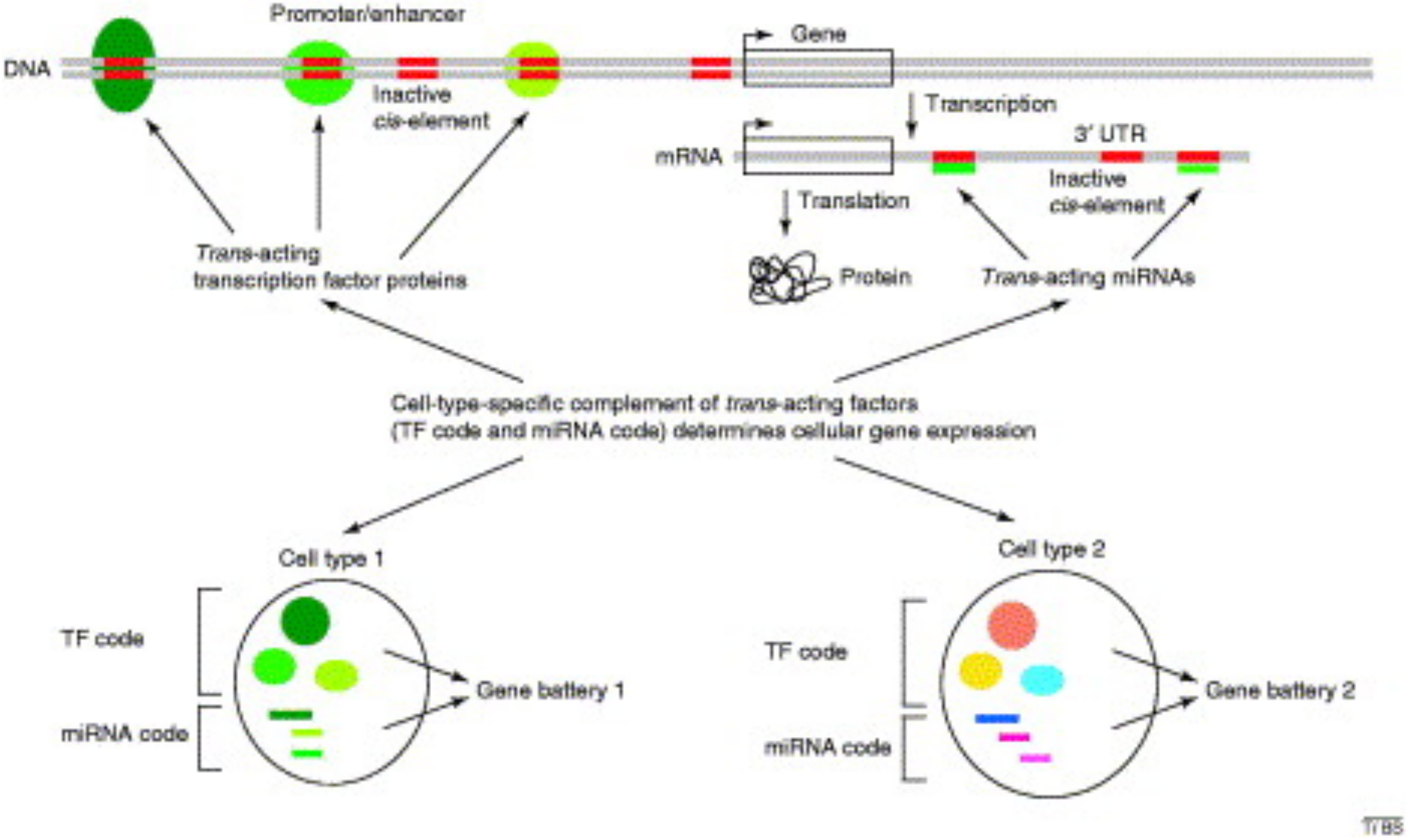


Jacek Krol, Inga Loedige & Witold Filipowicz
Nature Reviews Genetics **11**, **597-610 (September 2010)**
doi:10.1038/nrg2843

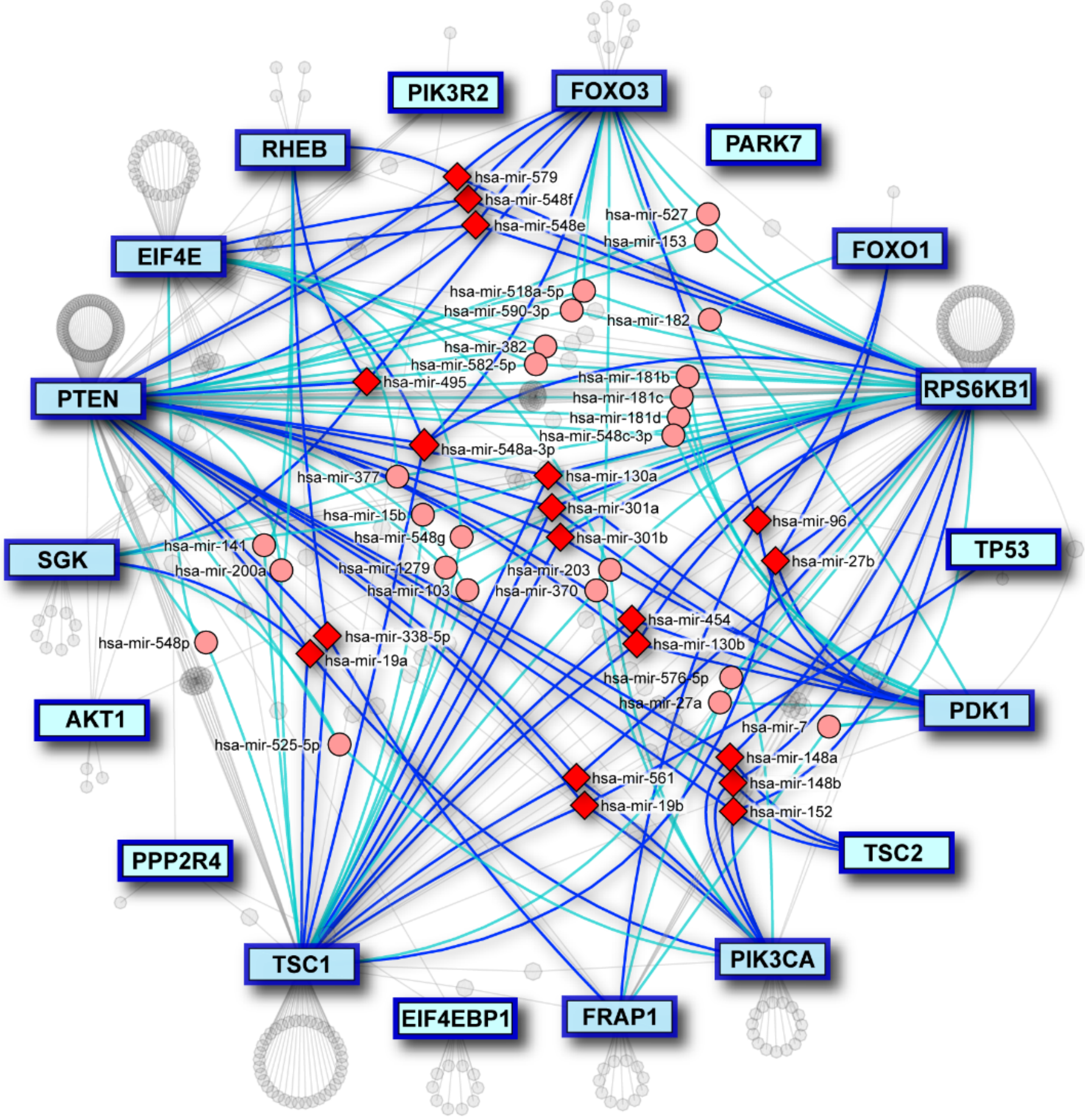
miRNA regulates mRNA



Common logic: transcription factors and miRNA



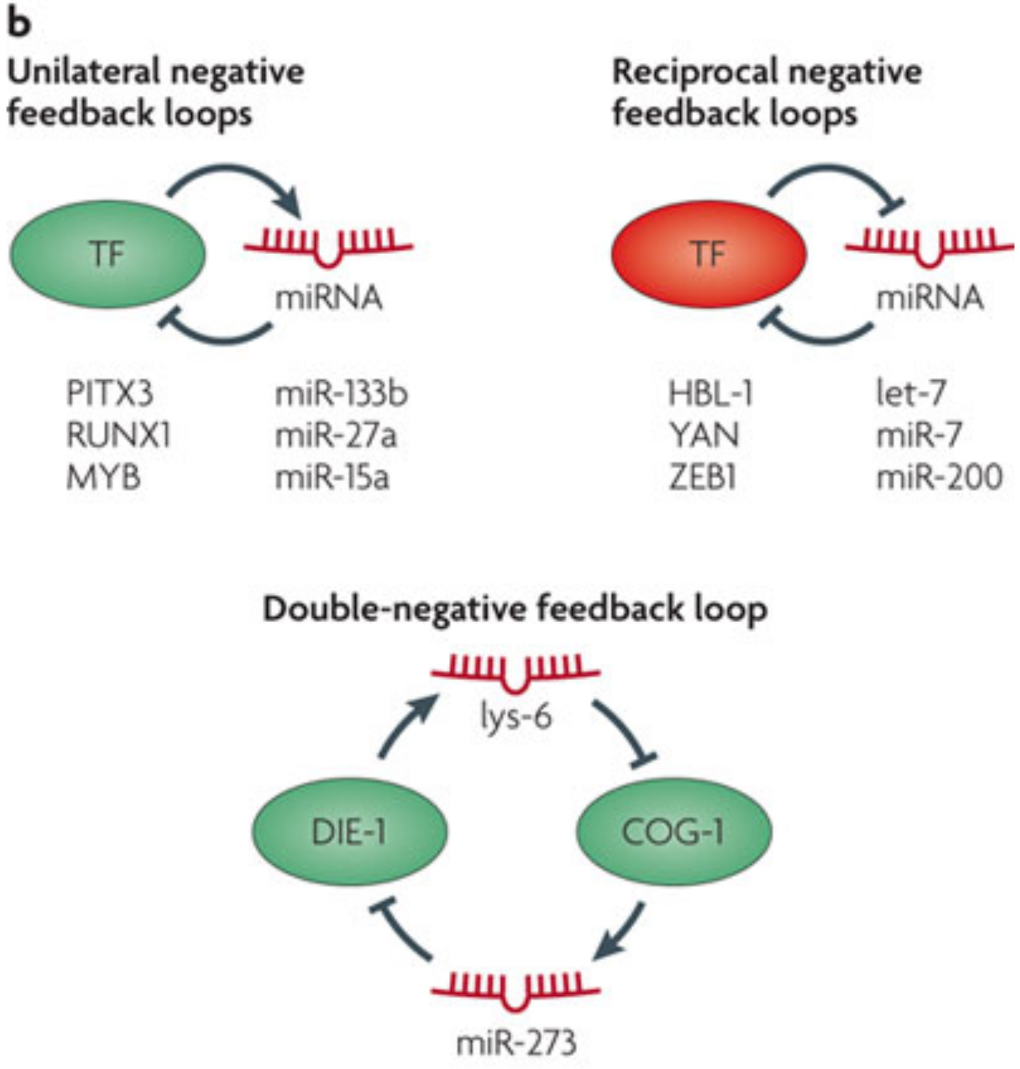
Hobert. Trends in Biochemical Sciences, Volume 29, Issue 9, 2004, 462 – 468. http://dx.doi.org/10.1016/j.tibs.2004.07.001



Regulation of oncogenes and tumor suppressors by miRNA

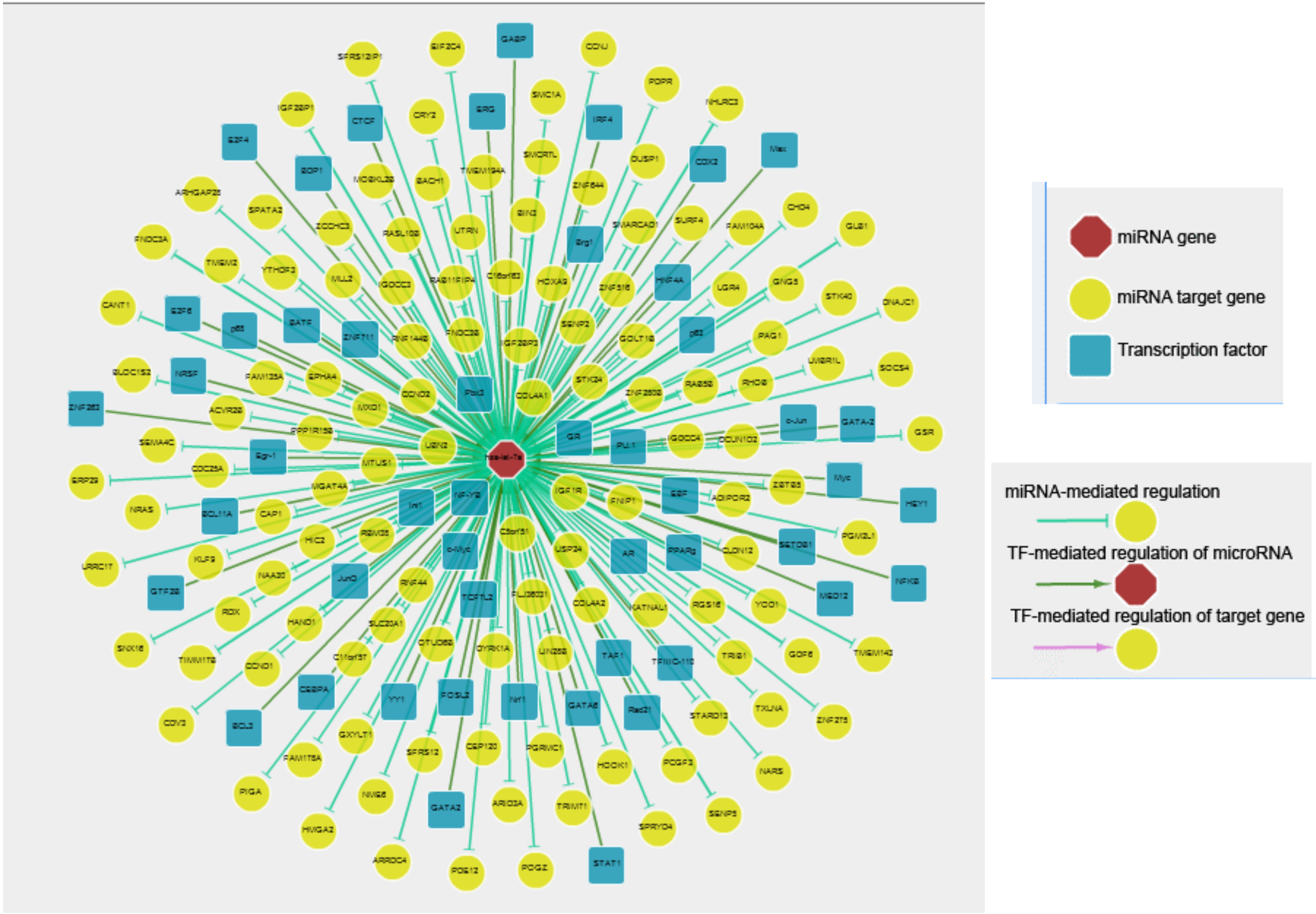
http://www.cs.toronto.edu/~juris/home.htm

Regulatory circuits

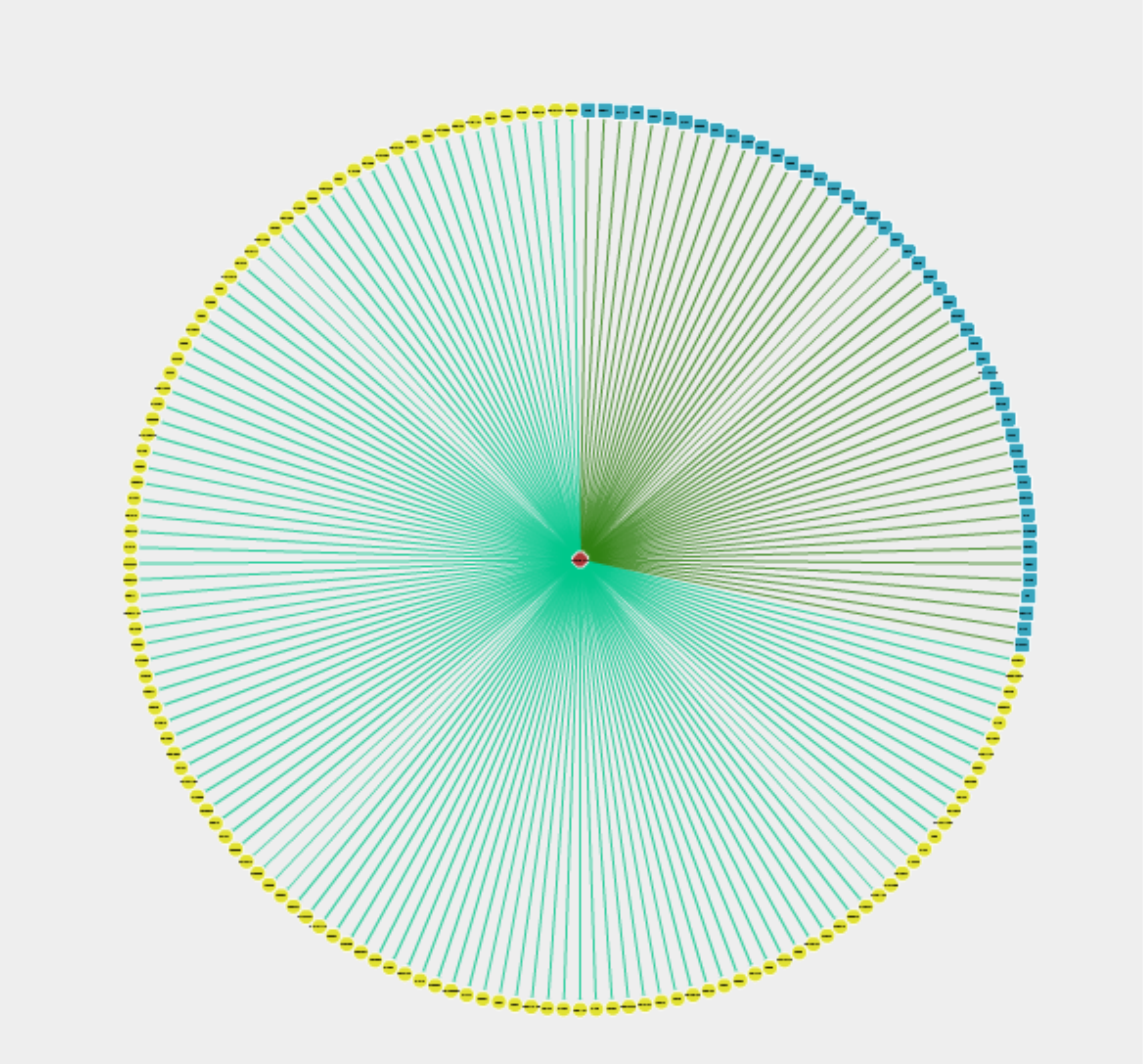


- Transcription factors can be + or –
- miRNA typically negative regulator
- Regulatory circuits typically contain both

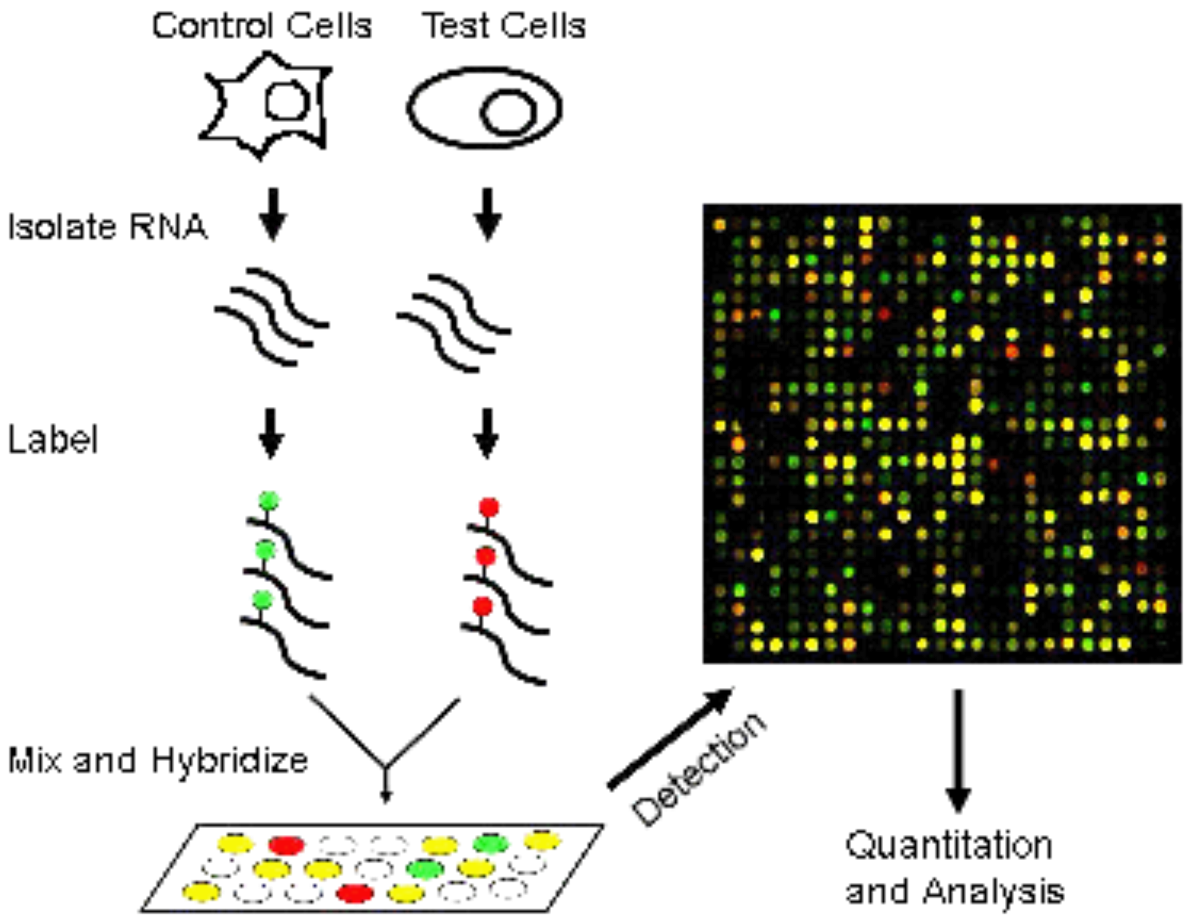
***Nature Reviews Genetics* 11, 597-610 (September 2010)**



http://deepbase.sysu.edu.cn/chipbase/tfmiRtargetNetworks

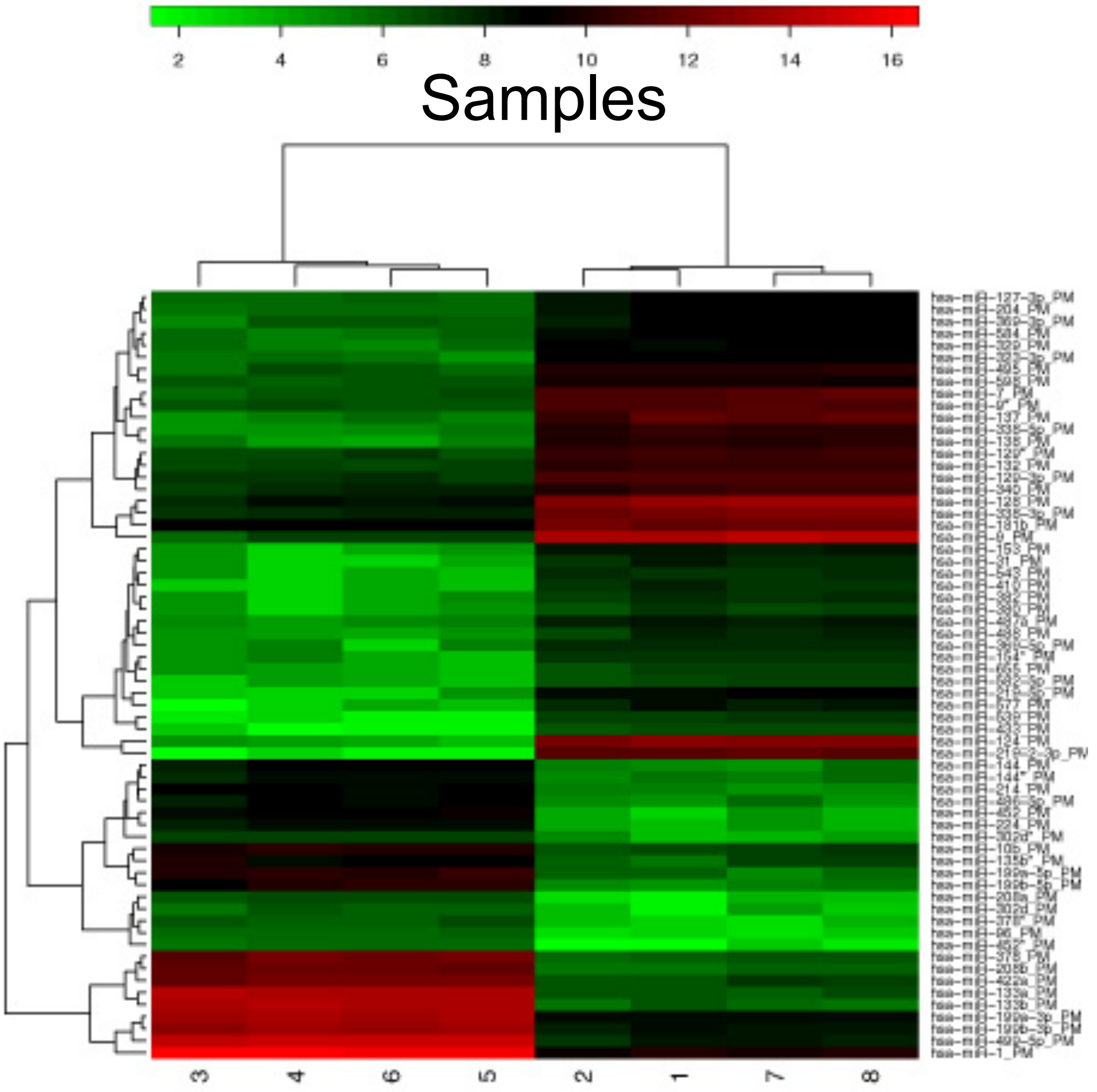


A microarray experiment



Data are variable!

- http://azcc.arizona.edu/research/shared-resources/gsr/services



Samples

Genes (mRNAs)

What are the trees telling you?

http://www.cbioc.com/en/services/bioinformatics-services/

Appendix: Databases and web sites

- DNA repositories
  - Genbank; NCBI (National Center for Biotechnology Information)
  - EMBL-bank; EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institute)
  - DDBJ (DNA Data Bank of Japan)

A few important websites

- NCBI
  - http://www.ncbi.nlm.nih.gov/
  - Multiple databases, tools
- EMBL-EBI
  - http://www.ebi.ac.uk/
  - Alignment tools in particular
- ExPASy (Expert Protein Analysis System – Swiss)
  - http://expasy.org/
  - Multiple tools, especially useful for secondary structure prediction

A sampling of NCBI Databases

- Nucleotides (Genbank entries)
- Gene (Refseq; annotated model organisms)
- Unigene (expression data)
- Homologene
- OMIM (Online Mendelian Inheritance in Man)
- Cn3D (3D structure info)
- CD (conserved domains)
- dbEST (expressed sequence tags)

Accession numbers vs. gi numbers

- Accession numbers
  - Unique, stable
  - AB123456.2 ← Version 2
  - New version made whenever any change is made to a sequence; old version info is retained
- gi (GenInfo) numbers
  - Assigned sequentially to each nucleotide sequence processed by Genbank
  - 12345678
  - New number whenever any change is made to a sequence; no relationship to old number
  - old number info is retained
- gi and Accession numbers are incremented in parallel

Refseq database

- Highly annotated entries from subset of organisms
- From the NCBI Glossary
  - RefSeq is the NCBI database of reference sequences; a curated, non-redundant set including genomic DNA contigs, mRNAs and proteins for known genes, and entire chromosomes
- Accession numbers start with two letters and an underscore
  - NM_123456.2 mRNA (version 2)
  - NP_123456.3 protein
  - NT_123456.4 contig
  - NC_000003.7 chromosome
  - XM_ or XP_ entries are 'models' predicted from sequence; no experimental evidence of existence

Some useful links

- NCBI Education pages
  - Glossary http://www.ncbi.nlm.nih.gov/books/NBK21106/
  - Tutorials http://www.ncbi.nlm.nih.gov/education/
- Tour of various NCBI databases and tools
- GenBank sample record explaining info in all fields
  - http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#ModificationsDateB
- Entrez nucleotide and protein FAQs
  - http://www.ncbi.nlm.nih.gov/books/NBK49541/
- The NCBI handbook
  - http://www.ncbi.nlm.nih.gov/books/NBK21101/