What is chemical biology?

The term “chemical biology” was published in 1930 by J. B. Leathes, (1930) “The birth of chemical biology”, in Lancet, 889–895. However, the usage of the term has changed dramatically with time, and the current definition cannot be traced back to a single source.

Here we will adopt the broad definition that **chemical biology is the application of chemical tools and ideas to biological problems**. It is interesting also to look at the leading journals in this field, all of which have appeared over the past few years. These include:

Nature Chemical Biology (est. 2005) http://www.nature.com/nchembio/index.html

"Nature Chemical Biology welcomes submissions from chemists who are applying the principles, language and **tools of chemistry to biological systems** and from biologists who are interested in understanding biological processes at the molecular level. The scope of the journal will emphasize four major themes of contemporary research in chemical biology: Chemical synthesis, Chemical Mechanisms in Biology, Expanding Biology through Chemistry, Expanding Chemistry through Biology.” (edited for length)

ACS Chemical Biology (est. 2006) http://pubs.acs.org/journals/acbcct/index.html

"ACS Chemical Biology is establishing a community approach **to foster interactions between scientists exploring cellular function from both a biological and chemical perspective**. Results will be published in which molecular reasoning has been used to probe questions through in vitro investigations, cell biological methods, or organismal studies. We welcome mechanistic studies on proteins, nucleic acids, sugars, lipids, and non-biological polymers. The journal serves a large scientific community, exploring cellular function from both chemical and biological perspectives.” **ACS Chemical Biology** is “a forum for the **publication of interdisciplinary research.**”

What is chemical biology?

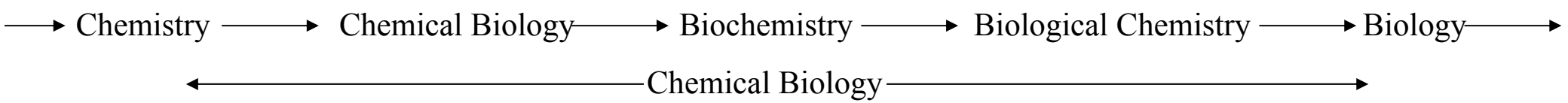
Chemistry and Biology (est. 1995) http://www.chembiol.com/

"Chemistry & Biology publishes reports of novel investigations in **all areas at the interface of chemistry and biology**. Chemistry & Biology strongly encourages submission of articles in which **chemical tools** are used to provide unique **insight into biological function and mechanism**.

ChemBioChem (est. 2006) http://www3.interscience.wiley.com/cgi-bin/jhome/72510898

"Contributions in ChemBioChem **cover chemical biology and biological chemistry**, bioinorganic and bioorganic chemistry, biochemistry, molecular and structural biology. That is, **research at the interface of chemistry and biology** that deals with the application of chemical methods to biological problems or uses life science tools to address questions in chemistry."

OK, so how is Chemical Biology different from Biochemistry? Here is one possible model:



→More than 50% of Nobel Prizes awarded in chemistry are for topics at the chemistry / biology interface:

2008: R. TSIEN, O. SHIMOMURA, and M. CHALFIE for the discovery and modification of *fluorescent proteins*.

2006: R. D. KORNBERG for his studies of the molecular basis of *eukaryotic transcription*.

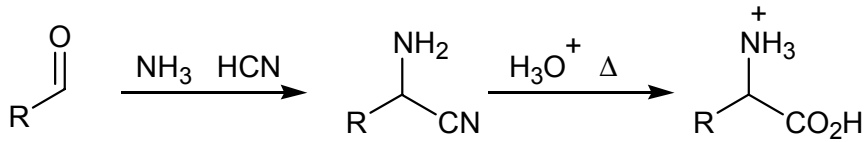
2004: A. CIECHANOVER, A. HERSHKO , and I. ROSE for the discovery of *ubiquitin-mediated protein degradation*

2003: P. AGRE, for the discovery of *water channels* and R. MACKINNON for *ion channels*.

Discovery of DNA

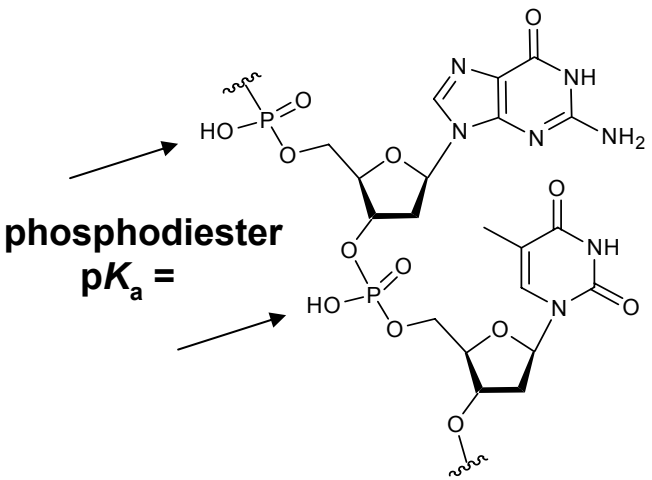
Friedrich Miescher (MD, born 1844 Basel)

Studied under organic chemist Adolf Strecker.



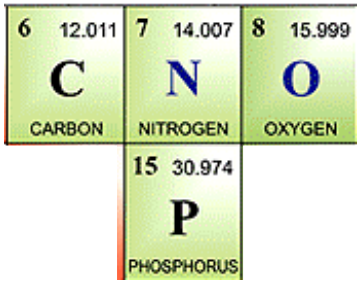
While at the Hoppe-Seyler Laboratory he subjected purified leucocyte nuclei from white blood cells to an alkaline extraction followed by acidification.

Miescher showed that the "nuclein" precipitate was a large molecule, acidic, resistant to pepsin, and rich in phosphorus.

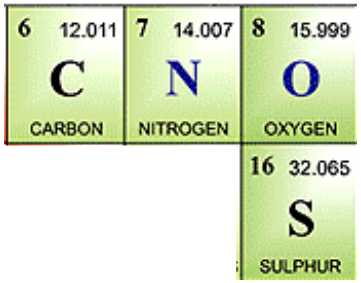


H_3PO_4: 2.15, 7.20, 12.35

"nuclein"



protein

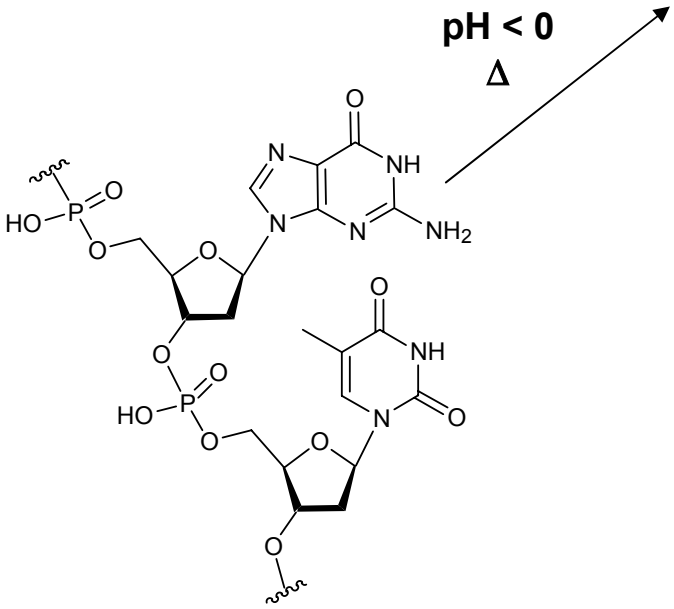


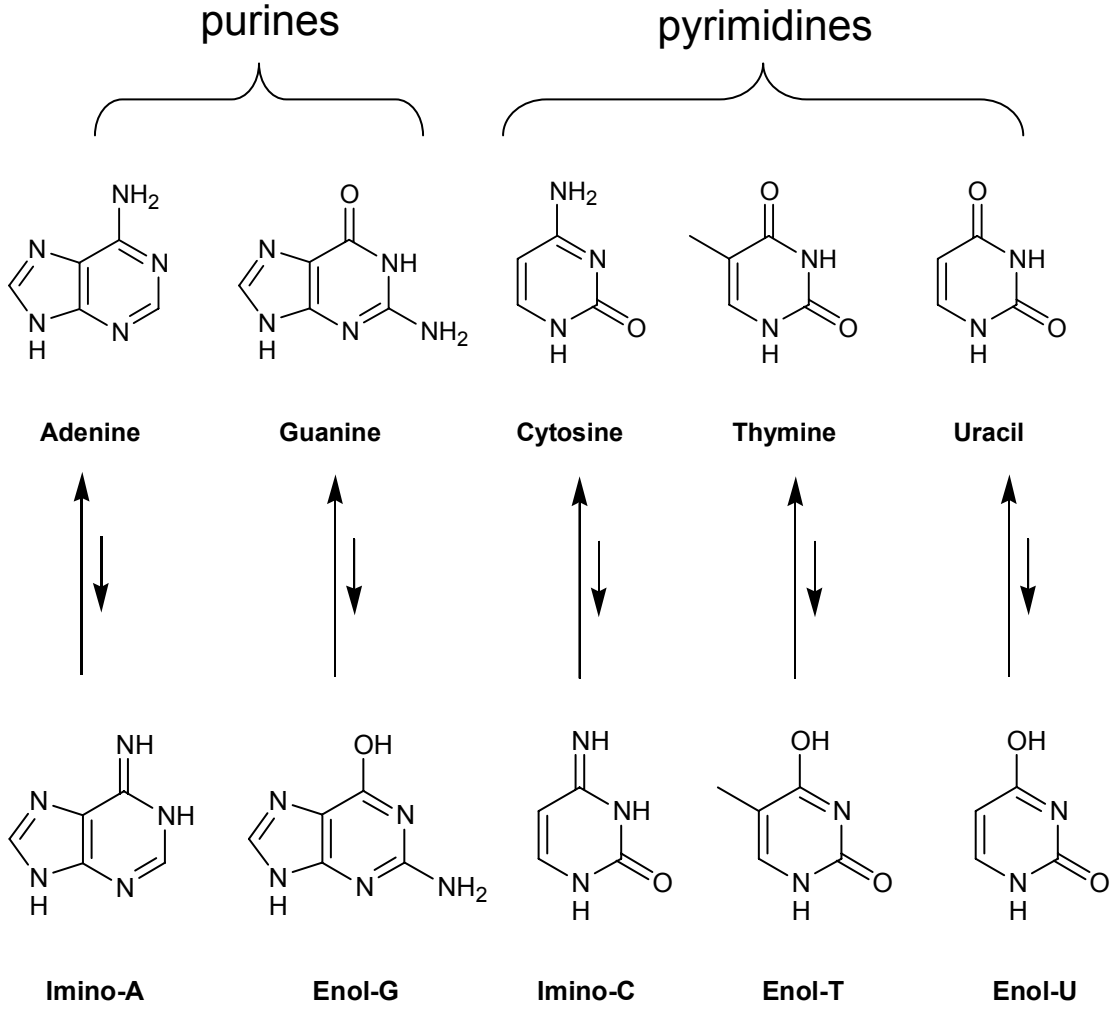
"....nuclein bodies will prove tantamount in importance to proteins"

Med.-Chem. Unters. ***1871****, 4, 441-460.*

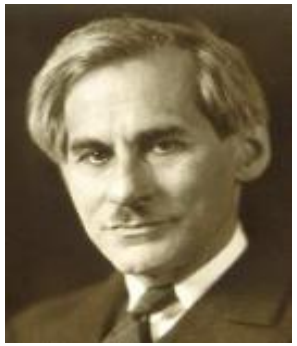
Composition of DNA

Albrecht Kossel: From **1879** to **1901** he and his students used hydrolysis and chromatography to isolate and characterize the five most abundant nucleobases from nuclein:



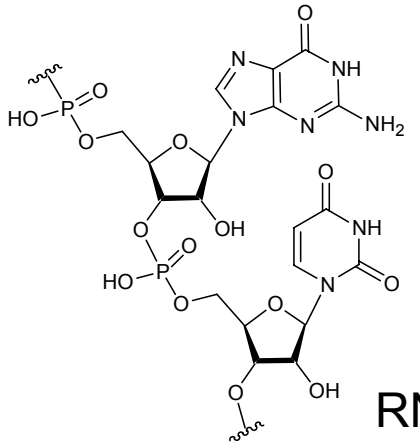


Composition of DNA

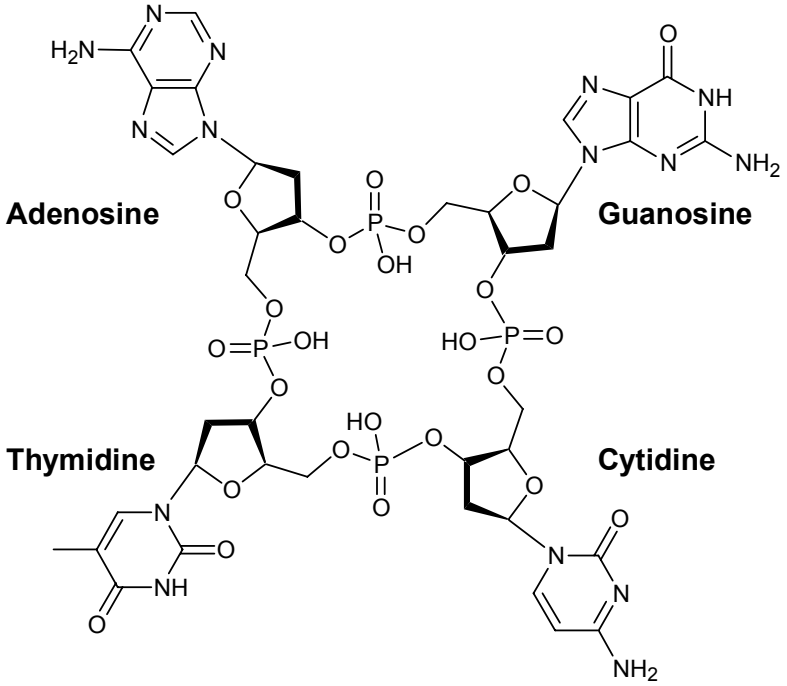


Phoebus Levene: Discovered ribose in **1909** and deoxyribose in **1929** and showed that the components were linked together in the order phosphate-sugar-base to form “nucleotide” units.

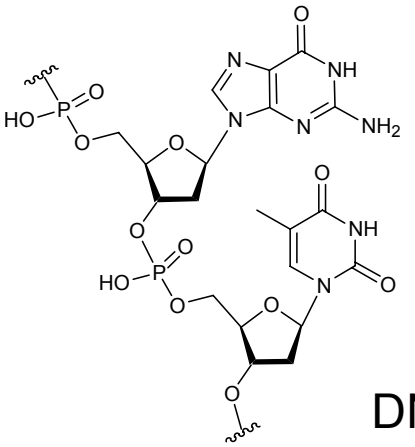
These discoveries lead to the terms ribonucleic acid (RNA) and deoxyribonucleic acid (DNA).



*J Biol Chem, **1919**, 40, 415–24*.

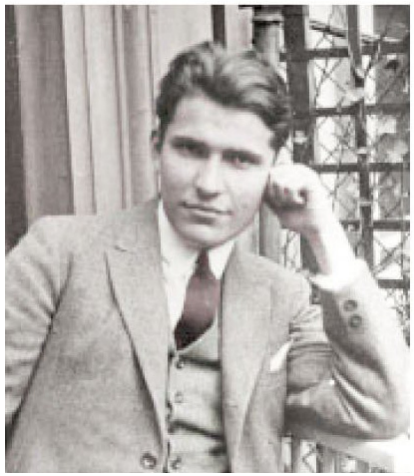


In Levene **1911** published the "tetranucleotide hypothesis“ which proposed that nucleic acids were made of equal amounts of adenine, guanine, cytidine, and thymidine.



*J Biol Chem, **1929**, 83, 793–802*.

Composition of DNA

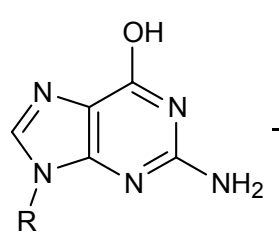


Erwin Chargaff: In **1950** reported that the amount of G = C, and that the amount of A = T.

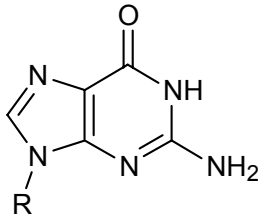
Organism	Base Composition (mole %)				Base Ratios		Ratio (A+T)/(G+C)
	A	G	T	C	A/T	G/C	
Human	30.9	19.9	29.4	19.8	1.05	1.00	1.52
Chicken	28.8	20.5	29.2	21.5	1.02	0.95	1.38
Yeast	31.3	18.7	32.9	17.1	0.95	1.09	1.79
Clostridium perfringens	36.9	14.0	36.3	12.8	1.01	1.09	2.70
Sarcina lutea	13.4	37.1	12.4	37.1	1.08	1.00	0.35

Base-Pairing Interactions

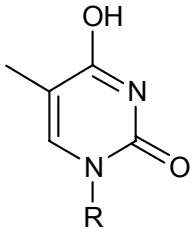
Jerry Donohue: In **1953** he informed Watson that, contrary to published work, the keto structures of guanine and thymidine were more important than the enol forms. Note that in the gas phase both forms are observed, but in solution the keto form of guanosine dominates by a factor of $10^4 - 10^5$.



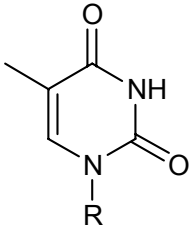
Guanosine (enol)



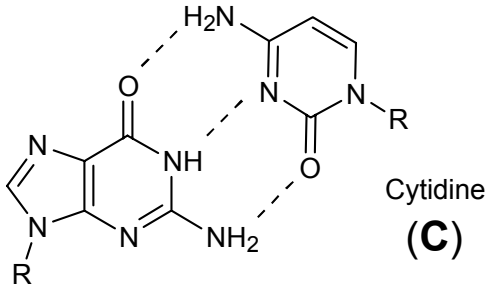
Guanosine (keto)



Thymidine (enol)

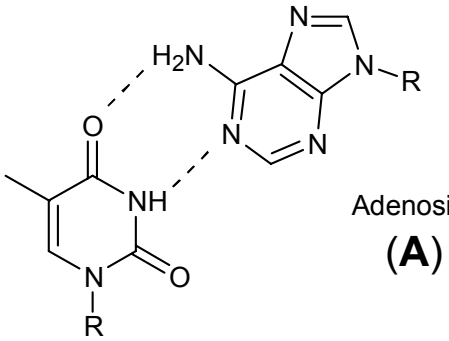


Thymidine (keto)



Cytidine
(**C**)

Guanosine
(**G**)

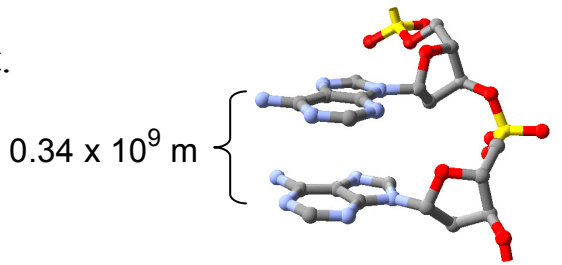


Adenosine
(**A**)

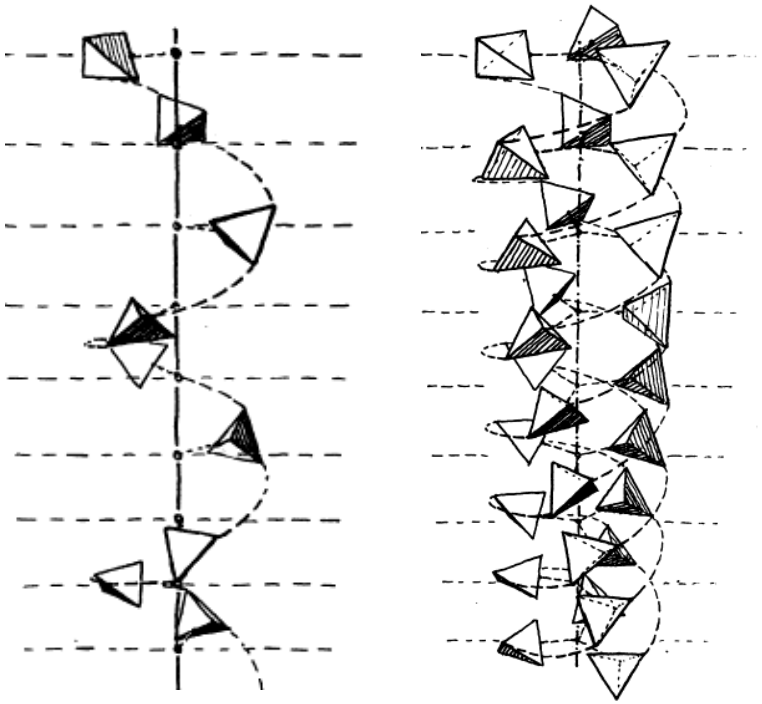
Thymidine
(**T**)

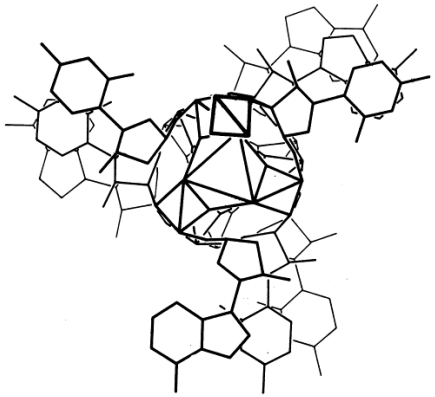
DNA Structure

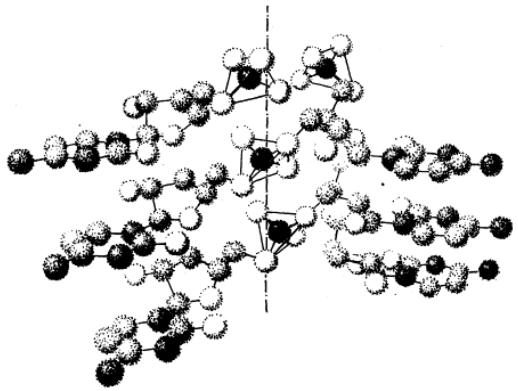
William Astbury: Using X-ray diffraction data, Astbury reported that DNA's structure repeated every ~30 Å and that the bases lay flat, stacked. 3.4 Å apart.



Linus Pauling: Published a triple helix DNA model using Astbury's X-ray diffraction data.







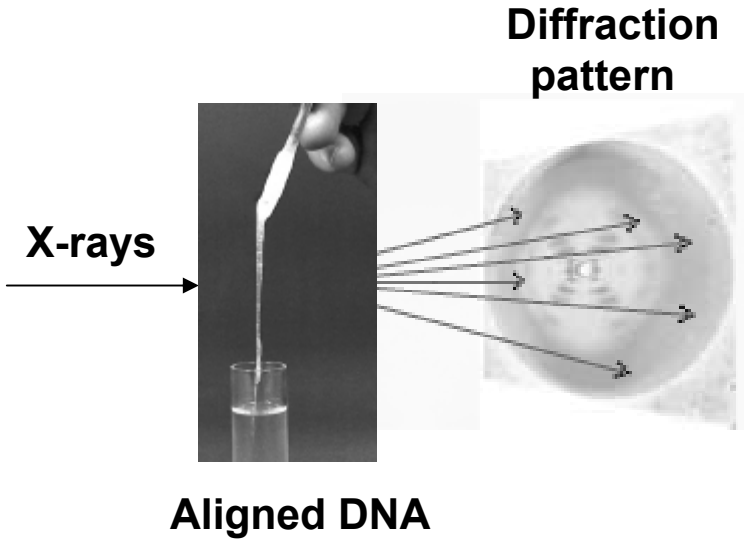
Proc Natl Acad Sci U S A. ***1953****, 39, 84-97.*

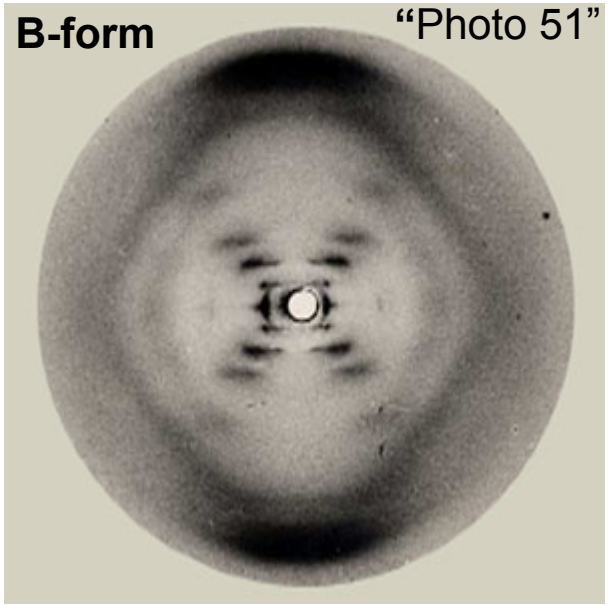
DNA Structure

Rudolf Signer:
Professor of organic chemistry at Bern University, provided Maurice Wilkins with high quality DNA.

Maurice Wilkins:
Started DNA diffraction studies with Ray Gosling in **1950** under Sir John Randal. Focused his work on B-form DNA.

Rosalind Franklin:
Obtained Signer's DNA from Wilkins in **1951**. Focused her work on A-form (low humidity) DNA.





Nature, 1953, 4356, 740-741.

DNA Structure

Rudolf Signer:
Professor of organic chemistry at Bern University, provided Maurice Wilkins with high quality DNA.

Maurice Wilkins:
Started DNA diffraction studies with Ray Gosling in **1950** under Sir John Randal. Focused his work on B-form DNA.

Rosalind Franklin:
Obtained Signer's DNA from Wilkins in **1951**. Focused her work on A-form (low humidity) DNA.

Maurice Wilkins showed "Photo 51" to James Watson, without Franklin's knowledge, in early **1953** or late **1952**.

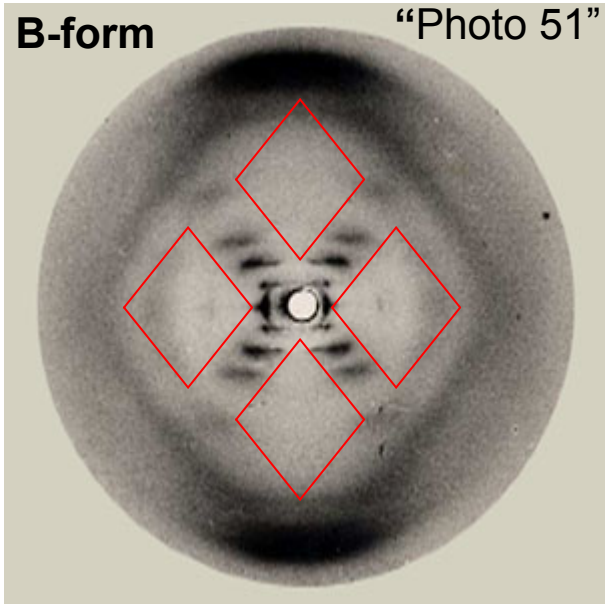
In February **1953** a written scientific report, including Franklin's interpretation of "Photo 51" was given to Francis Crick by his thesis supervisor **Max Perutz**.

*EMBO reports, 2003, **4**, 11, 1025–1026.*

"We were not aware of the details of the results presented there (from King's College) when we devised our structure....

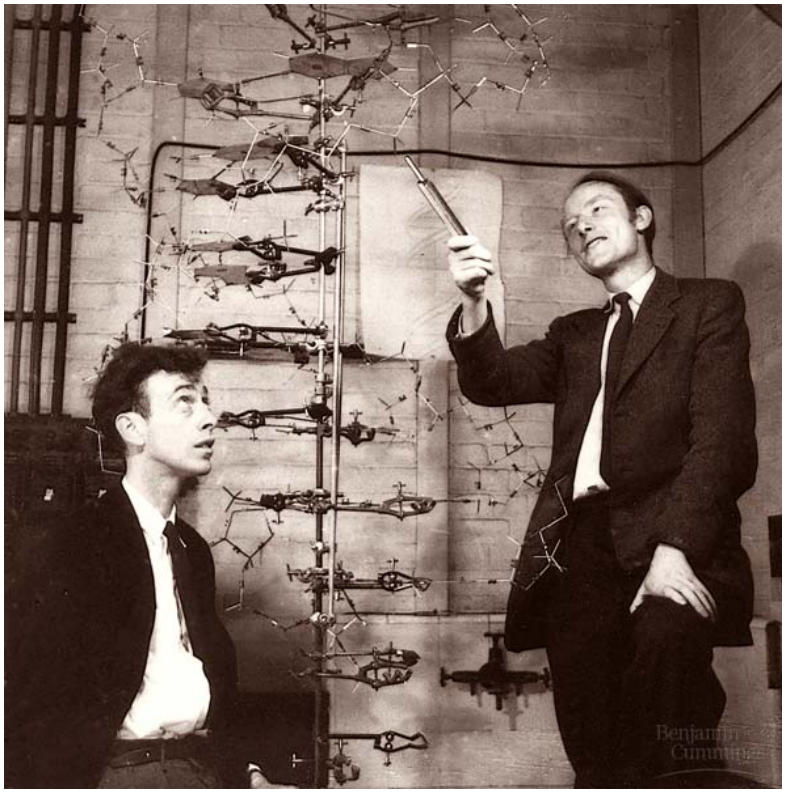
We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. Wilkins and Dr. Franklin and their co-workers."

*Nature, **1953**, 4356, 737-738.*



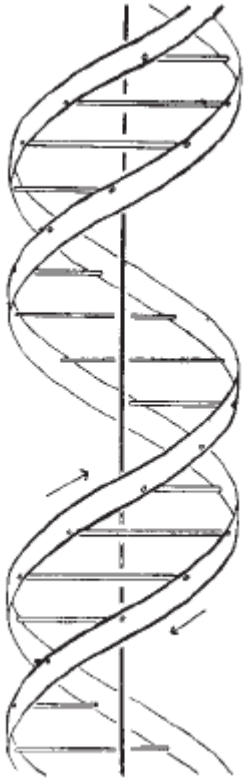
→vertical axis, bases inside

DNA Structure



James Watson & Francis Crick in **1953** published their model for the B-form double helix.

Received Nobel Prize for Medicine or Physiology, along with Maurice Wilkins, in **1962**.

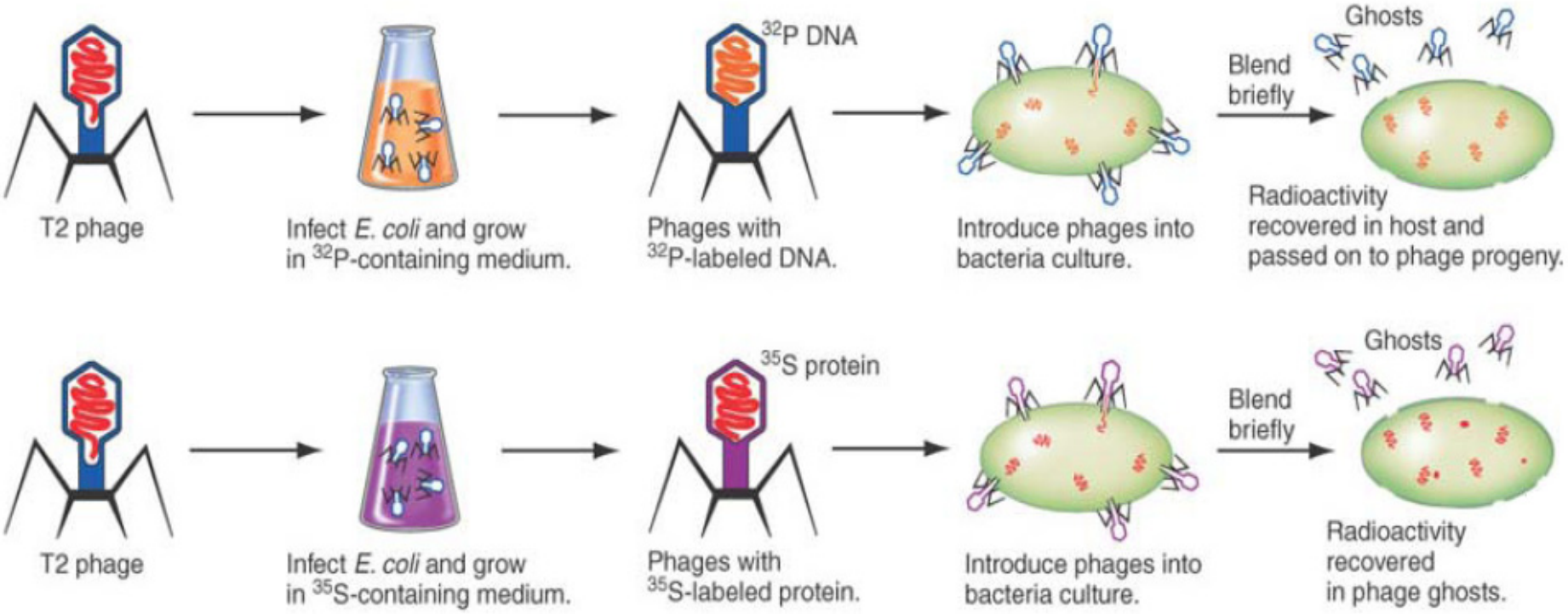


*Nature, **1953**, 4356, 737-738.*

"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."

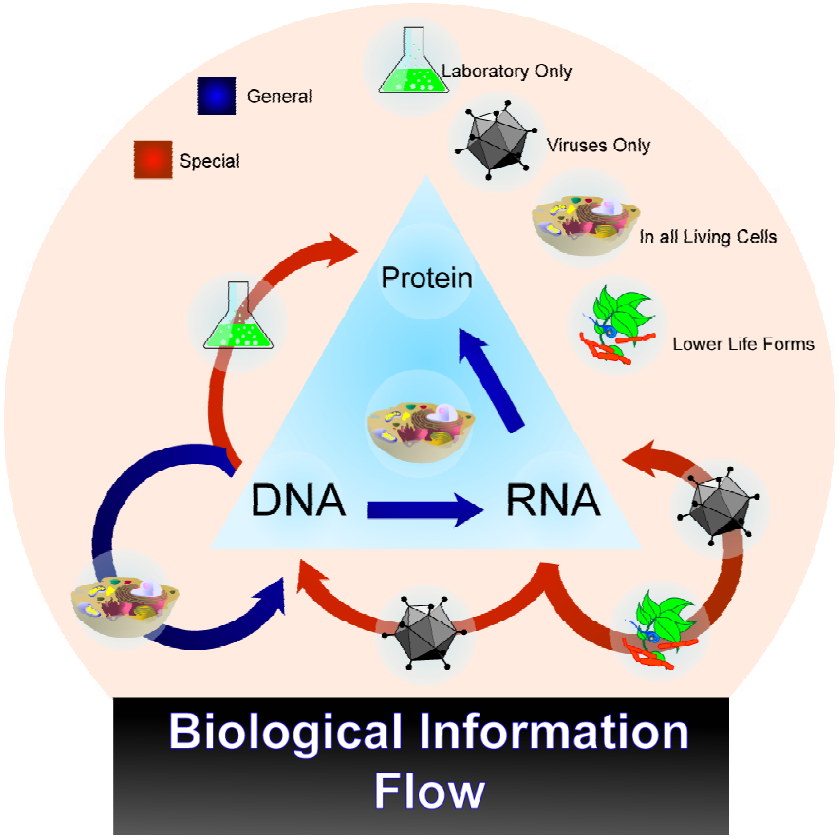
DNA Function

Alfred Hershey and **Martha Chase:** reported their "blender experiment" in **1953**, and DNA was finally accepted as the molecule of inheritance.



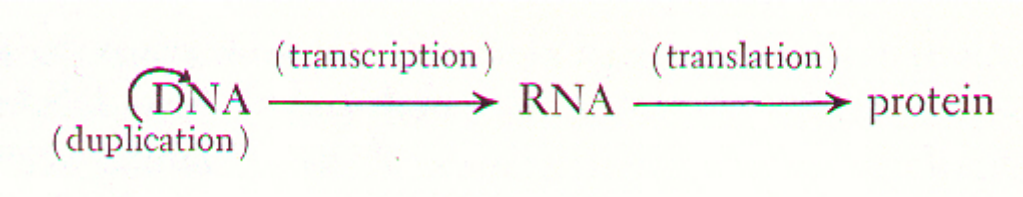
The Central Dogma of Molecular Biology

One Current Model (2008):

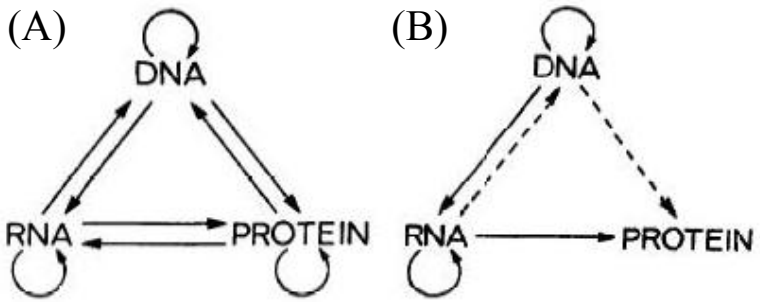


http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

J. Watson's Model (1965):



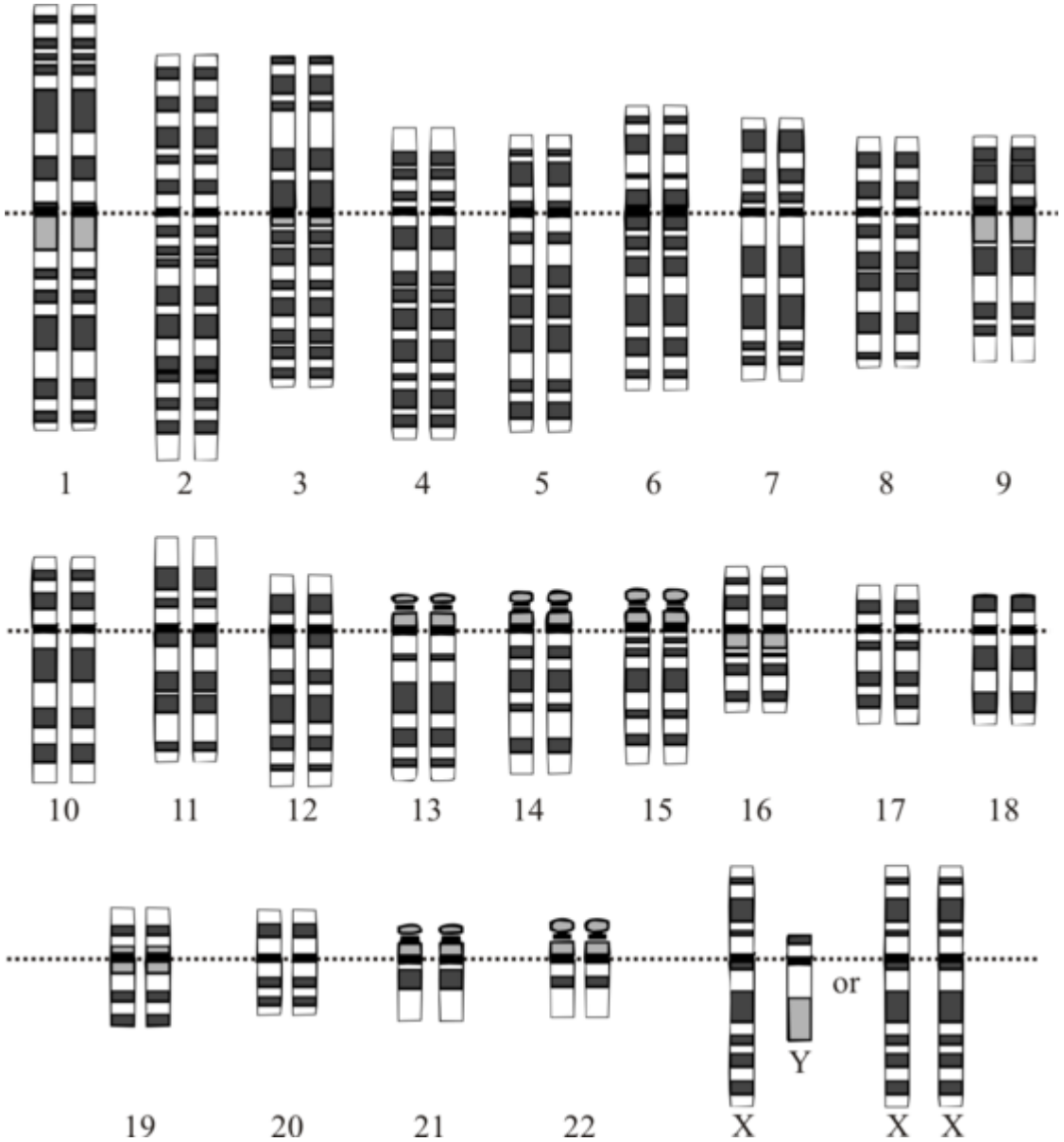
F. Crick's Model (1958):



Models for the flow of genetic information illustrating all possible transfers of (A) and those that are permitted (B). Crick concluded that once information was transferred from nucleic acid (DNA or RNA) to protein it could not flow back to nucleic acids.

Many exceptions to Watson's (simplified) model are known: non-coding RNA (tRNA, rRNA, etc.), reverse transcription (RNA viruses), "junk" DNA, epigenetics, prions, trans-splicing, transposons, gene rearrangements (information transfer that is independent of duplication), and the one-gene-one-protein hypothesis.

Human Genome



Human Genome Project
Projected cost ≈ 4 x 10^9 CHF

Craig Venter & Celera Genomics:
Final cost ≈ 4 x 10^8 CHF

85% complete in 2001.
92% complete in 2005.

Sequences can not be patent protected.

Near-term cost projections ≈ 1'000 CHF per human genome

Genome vs. Proteome

Species	Genome size (Mb)	Estimated Number of genes
Mycoplasma genitalium (bacterium)	0.58	500
Streptococcus pneumoniae (bacterium)	2.2	2300
Escherichia coli (bacterium)	4.6	4400
Saccharomyces cerevisiae (yeast)	12	5800
Arabidopsis thaliana (plant)	125	25,500
Caenorhabditis elegans (worm)	97	19,000
Drosophila melanogaster (fruit fly)	180	13,700
Mus musculus (mouse)	2500	29,000
Homo sapiens (human)	2900	27,000
Oryza sativa (rice plant)	466	45-55,000

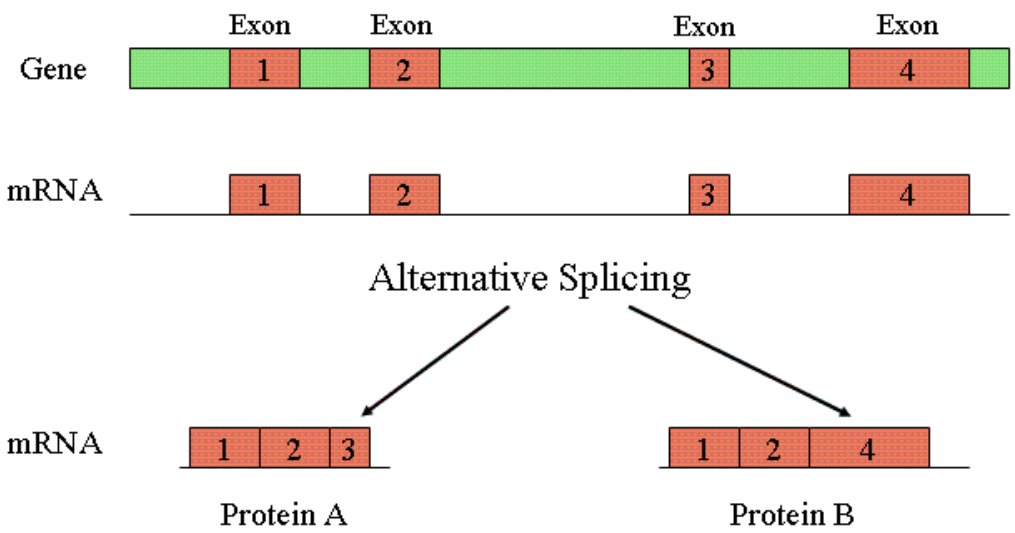
According to Molecular Biology of the Gene, 5th ed., Peason Benjamin Cummings (Cold Spring Harbor Laboratory Press, 2005).

What is the "gene density" (in base pairs per gene) of *H. sapiens* compared to *C. elegans*?

Why does the number of genes estimated from genomic DNA sequences underestimate the number proteins?

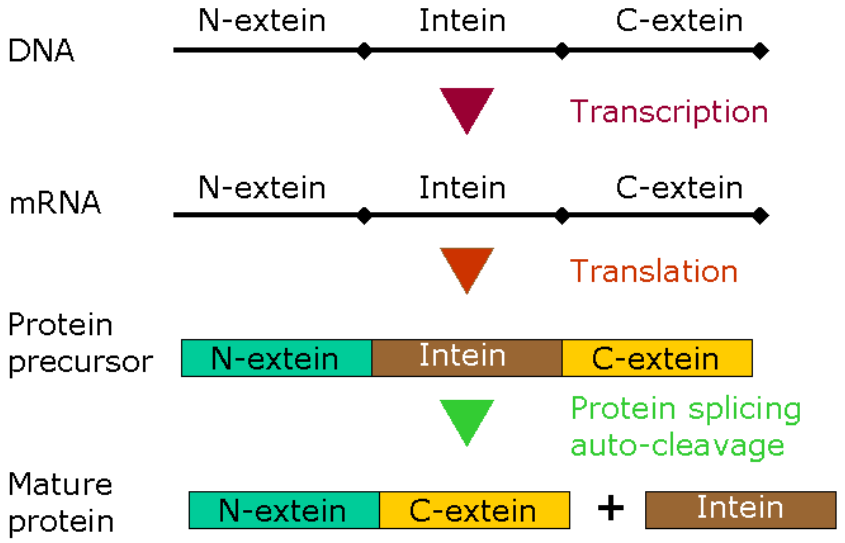
Genome vs. Proteome

Alternative Splicing of RNA



This is called alternative splicing, and can produce different forms of a protein from the same gene. The different forms of the mRNA are called "transcript variants," "splice variants," or "isoforms." The current record-holder for alternative splicing is a Drosophila gene called Dscam which has ~38'000 splice variants. This gene has 95 alternate exons and encodes an axon guidance receptor.

Splicing of Protein*



An intein is the protein equivalent of intron. A self-splicing intein catalyzes its own removal from the host protein through a posttranslational process of protein splicing. A mobile intein displays a site-specific endonuclease activity that confers genetic mobility to the intein through intein homing. Inteins can evolve into new structures and new functions, such as split inteins that do trans-splicing.

Annu. Rev. Genet. 2000. 34:61–76

* = to be covered in this course

Genome vs. Proteome

Known post-translational protein modifications involving the addition or modification of functional groups:

Acetylation*: the addition of an acetyl group, usually at the N-terminus of the protein

Methylation: the addition of a methyl group, usually at lysine or arginine residues.

Biotinylation*: acylation of conserved lysine residues with a biotin appendage

Glutamylation: covalent linkage of glutamic acid residues to tubulin and some other proteins

Glycylation: covalent linkage of one to more than 40 glycine residues to the tubulin C-terminal tail

Glycosylation*: the addition of a glycosyl group to either asparagine, hydroxylysine, serine, or threonine, resulting in a glycoprotein

Isoprenylation: the addition of an isoprenoid group (e.g. farnesol and geranylgeraniol)

Lipoylation: attachment of a lipoate functionality

Phosphopantetheinylation*: the addition of a 4'-phosphopantetheinyl (ppan) moiety from coenzyme A, as in fatty acid, polyketide, non-ribosomal peptide and leucine biosynthesis

Phosphorylation*: the addition of a phosphate group, usually to serine, tyrosine, threonine or histidine

Sulfation: the addition of a sulfate group to a tyrosine

Selenation: the exchange of a sulfur group with selenium

Citrullination: or deimination, the conversion of arginine to citrulline

Deamidation: the conversion of glutamine to glutamic acid or asparagine to aspartic acid

Disulfide bridges*: the covalent linkage of two cysteine amino acids

C-terminal amidation*

* = to be covered in this course

Important Challenges to Proteomics

1. Which proteins are encoded by the genome and under what conditions?
2. What are the sequences and post-translational modifications of each protein?
3. What does each protein do?
   (a) Is it an enzyme, a structural scaffold, a mediator of signal transduction, an antibiotic, a hormone?
   (b) To which molecules does each protein bind?
   (c) How does over-expression or under-expression of each protein affect the cell and the expression of other genes?
   (d) Where in the cell is the protein localized?
   (e) How is its activity regulated?

The study of natural proteins and the creation of non-natural ones require the ability to produce and manipulate proteins. The production of proteins using biochemical methods (recombinant DNA technology) can, in favorable cases, provide access to large amounts of protein and allow for site-specific mutations. However, the natural genetic code is typically limited (not always) to the 21 natural proteinogenic amino acids. To probe the questions above, other types of modifications are often required like the site-specific incorporation non-proteinogenic amino acids, phosphorylated residues, novel glycosylation patterns, fluorescent labels, photochemical triggers, lipid anchors, etc. Over the past 10 years a number of powerful methodologies have been developed for producing proteins with such modified groups. The first part of Chemistry 434 will focus on these methodologies.

Some Relationships Between Analytical and Preparative Chemical Tools

Chemical tools

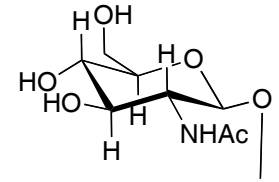
Analytical	**Preparative**
Protein sequencing.	Protein synthesis.*
Identification of post-translational modifications.	Site-specific introduction of post-translational modifications.*
Protein conformation and structure.*	Chemical cross-linking and rational design/modification.*
Protein localization, and catalysis.	Site-specific protein labeling with spectroscopic probes.*
Identification of protein-protein, protein-DNA, protein-RNA, protein-small molecule binding interactions.*	Protein-capture through binding interactions, and the preparation of mutants to probe the contributions of individual residues for binding.*

* = to be covered in this course

Examples of Problems in Preparative Chemistry

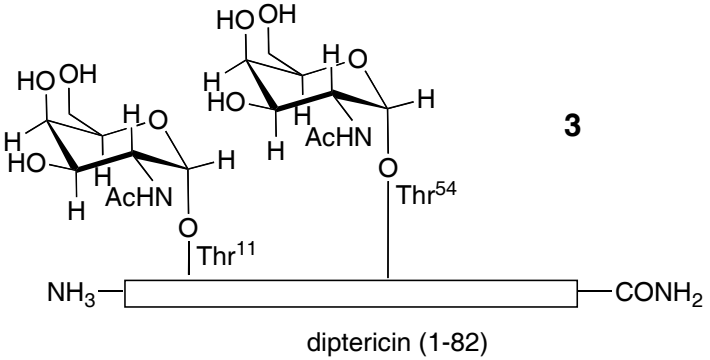
For example: how would you make the 36-residue peptide T20 (**1**) (also called Fuzeon, a new anti-HIV drug from Roche); or the glycopeptide **2**, which is a model of the repeating C-terminal domain of RNA polymerase II, in order to study the structure and function of this enzyme; or the antimicrobial glycoprotein diptericin (**3**) isolated from insects, to study the mechanism of its antimicrobial action; or an analogue of ribonuclease A containing a 5,5-dimethylproline (**4**) residue in place of proline, to examine the effects on folding; or a modified form of a kinase containing a caged photoactivatable phospho-group (**5**), in order to use time-resolved techniques to study the function of the phosphoprotein in cell migration; or a lipidated version of green fluorescent protein (**6**) to study its mobility in membranes; or replace site specifically an amino acid in a protein for a p-benzoyl-L-phenylalanine residue (**7**) to allow photo-crosslinking ???

Ac-YTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWF-NH_2 **1**



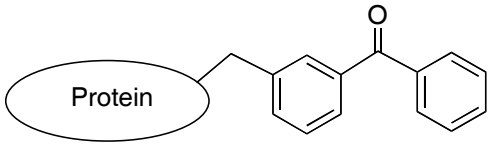
2

Ac-Ser-Tyr-Ser-Pro-Thr-Ser-Pro-Ser-Tyr-Ser-NH_2

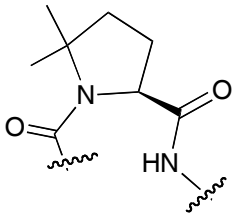


3

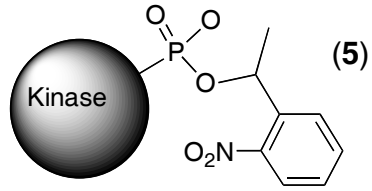
diptericin (1-82)



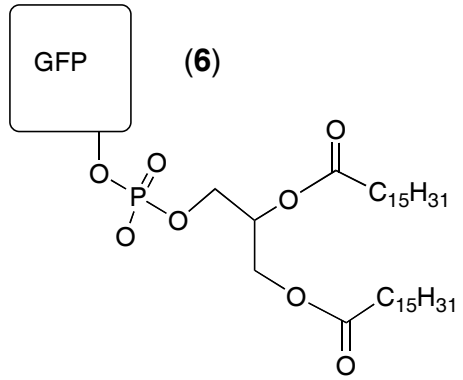
(**7**)



5,5-dimethylproline **(4)**



(**5**)



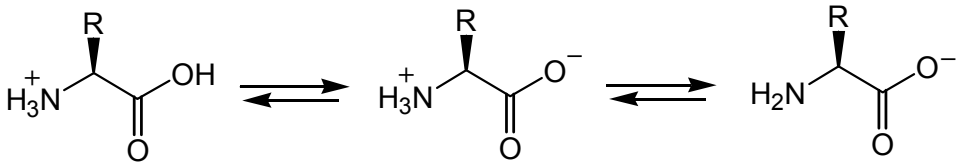
(**6**)

Back to the Basics: Amino Acids, Peptide Bonds, and Proteins

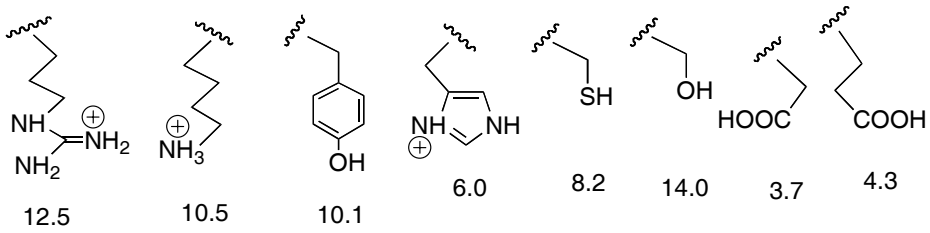
Proteinogenic amino acids

There are over 500 naturally occurring α-amino acids (ignoring β-, γ- and δ-amino acids), but the genetic code in nearly all organisms encode for only 21 different α-amino acids. All of these except glycine have a chiral center at the α-position, which has the *S* absolute configuration. The exception is cysteine, which according to the CIP-nomenclature system is *R*. Two others (*2S,3R*-threonine and *2S,3S*-isoleucine) have chiral centers at the ß-position. All the proteinogenic amino acids are classified as L-amino acids. Bacteria also produce D-amino acids, which are needed amongst other things, for the biosynthesis of the cell wall peptidoglycan.

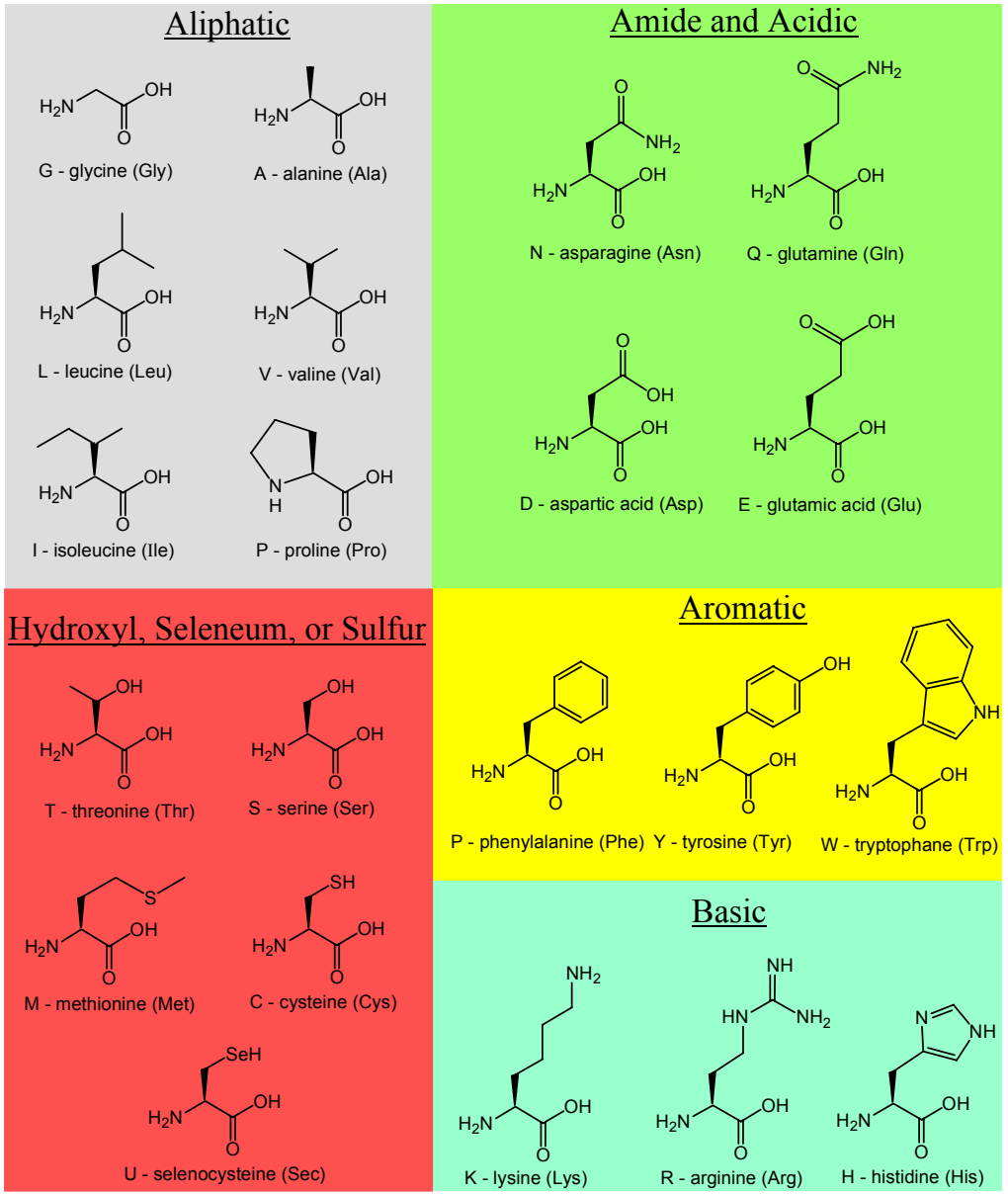
Amino acids are zwitterionic in neutral aqueous solution:



The side chains (R) of many amino acids contain charged groups:



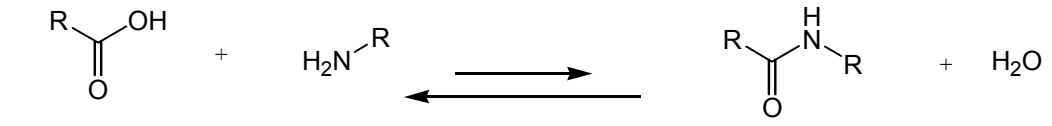
These pK_a values apply only in free aqueous solution. In Other environments, the pK_a values may change dramatically.



Amide Bonds

Formal Dehydration Reaction:

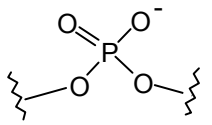
Amides are synthesized from carboxylic acids and amines. This reaction formally involves the loss of a water molecule.



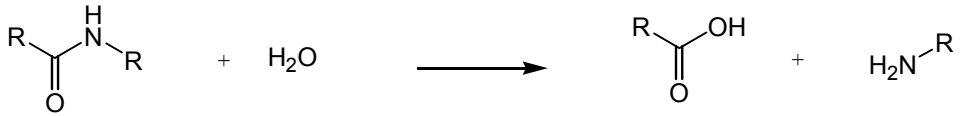
Carboxylic Acid Amine Amide Water

Reverse Reaction:

The reverse reaction is extremely slow in neutral water at RT, with an estimated half life of 7 years. For phosphodiester bonds, estimated half life = 1'000 years (RNA), 200 x 10^6 years (DNA):



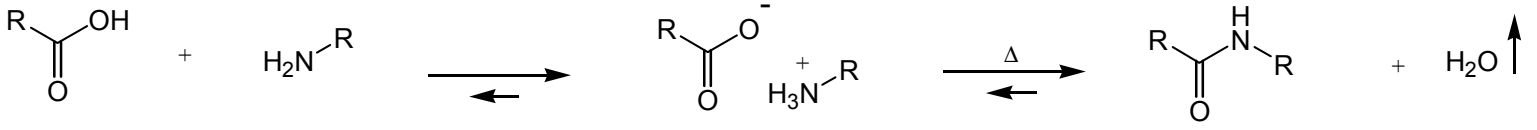
Phosphodiester



Amide Water Carboxylic Acid Amine

Synthesis:

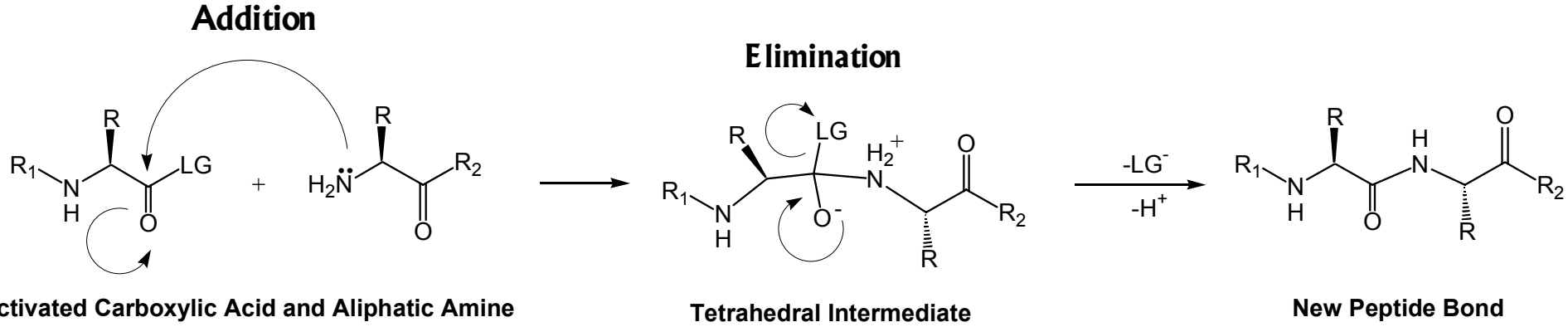
It is possible to use dehydration for amide bond synthesis, but it requires very high temperatures (200 ^oC) and does not work for complex "R" groups (like most amino acids).



Carboxylic Acid Amine Salt Amide Water

Peptide Bonds

Both chemical and biosyntheic peptide bonds are synthesized from "activated" carboxylic acids and an amine nucleophile:

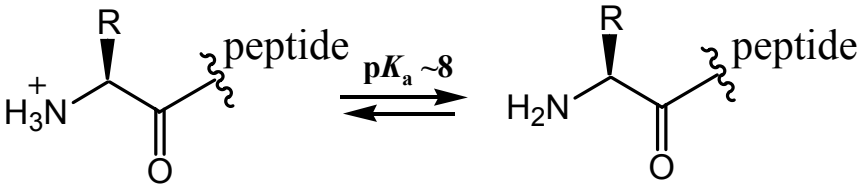


Biosynthesis: **LG** = tRNA or ppan; **R$_1$** = polypeptide chain or H; **R$_2$** = tRNA or ppan

Growing chain formed in N → C terminal direction

Chemical Synthesis: **LG** = chloride, succinate, HOBt, urea, etc.; **R$_1$** = protecting group; **R$_2$** = polypeptide chain and linker to a solid-support

Growing chain formed in C → N terminal direction



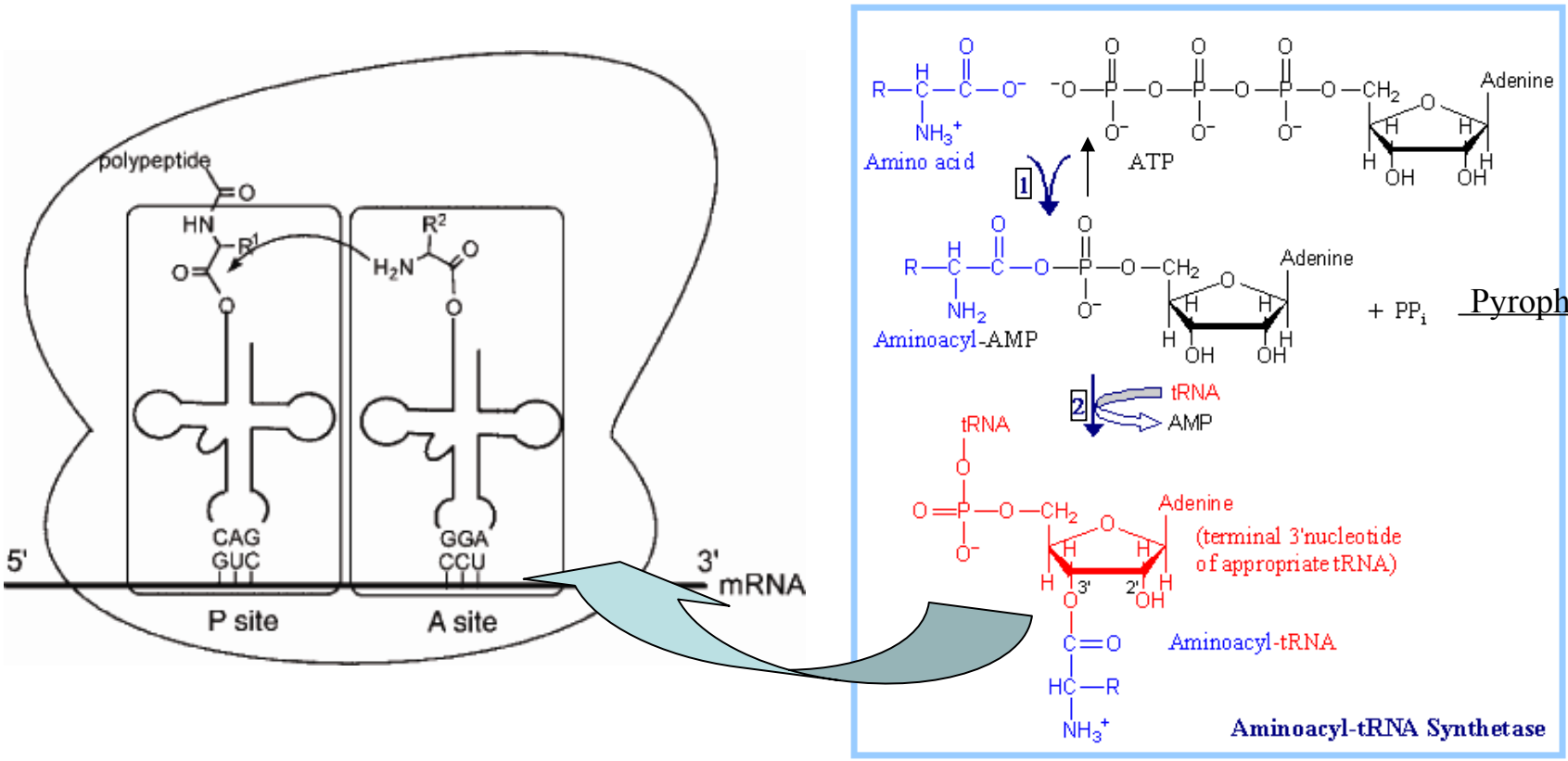
→How is this pK_a value different than free amino acids?

→At pH = 7.0 what is the fraction of the ammonium ($-NH_3^+$) versus amine ($-NH_2$) form?

Henderson-Hasselbach equation: $pH = pK_a + \log([A] / [HA])$

Biosynthesis of Polypeptides

Biosynthesis:
Amino acids are converted into esters by the ATP-dependent catalysis of Aminoacyl-tRNA Synthetase in a two-step process. Each amino acid requires a different Aminoacyl-tRNA Synthetase. Peptide bonds are formed in a portion of the ribosome that contains only RNA.

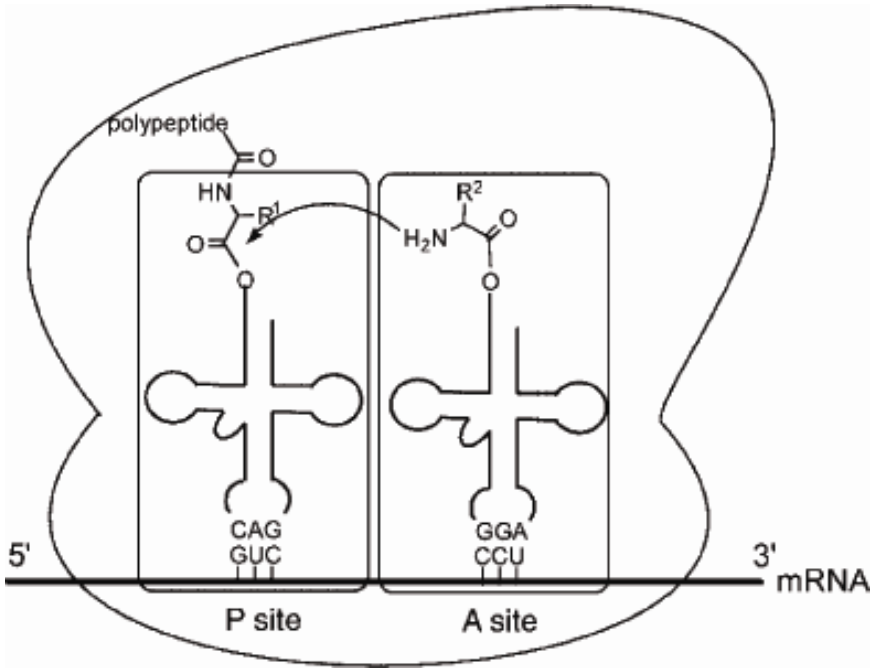


1 Reversible. Driven by subsequent PPi hydrolysis.

2 This is one of two key steps needed for accurate translation of mRNA.

99.98% accurate! How can small differences is codon-anticodon base pairing lead to such high fidelity?

The Ribosome and Translation



At least two proofreading mechanisms exist to prevent errors:

1. Proofreading of aminoacyl-AMP and tRNA pair.

2. Kinetic proofreading before peptide bond formation.

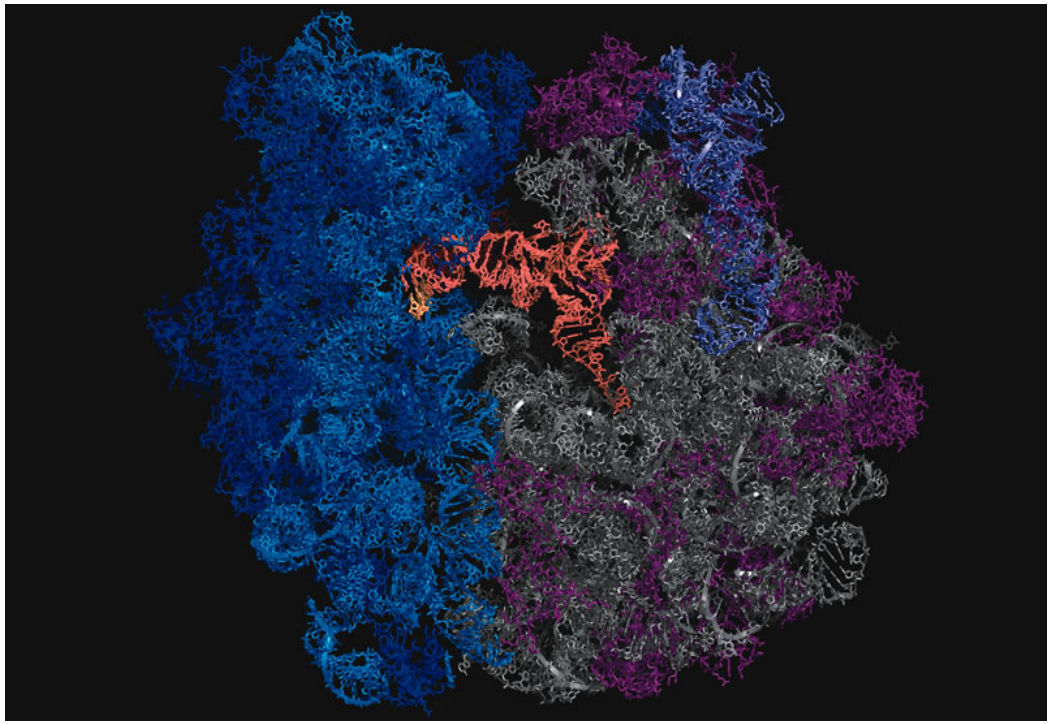
A delay is introduced between the binding of an aminoacyl-tRNA to the codon and the formation of the peptide bond to allow errors to be corrected:

EF-Tu-GTP binds an aminoacyl-tRNA and bring it into the A-site. EF-Tu allows the anticodon to interact with the codon but prevents peptide bond formation. An incorrect tRNA will bind weakly to the codon and will dissociate from the codon before an incorrect amino acid is incorporated into the polypeptide. Correct codon-anticodon matching triggers hydrolysis of GTP by the EF-Tu, after which EF-Tu-GDP dissociates.
Peptide bond formation proceeds.

See movies:

http://pubs.acs.org/isubscribe/journals/cen/85/i08/html/8508cover.html

The Ribosome and Translation



"If genomic DNA is the cell's planning authority, then the ribosome is its factory, churning out the proteins of life".

The ribosome found in the bacterium Escherichia coli is made up of three RNA components and more than 50 proteins. It weighs about **2.5 million daltons**. Eukaryotic versions have four RNAs and about 80 proteins and weigh about 4 MDa.

Since the average weight of an amino acid is ~100 daltons, this is equivalent to a 80 kg person working with an object weighing 3 grams.

What is the rate of translation?

The rate of translation in *E. coli* is about 15 amino acids per second.

Chemical synthesis is ~ 1 amino acid per hour.

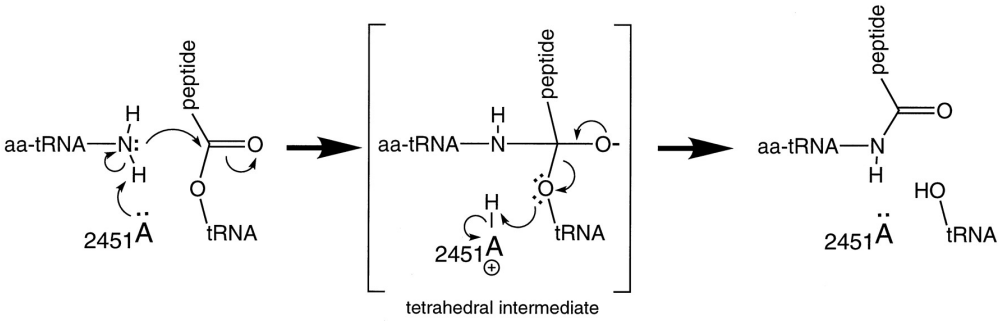
What is the accuracy (fidelity) of translation?

For *E. coli* : 99.2 – 99.98% accurate!

The yield for a typical chemical synthesis is ~96 % per amino acid.

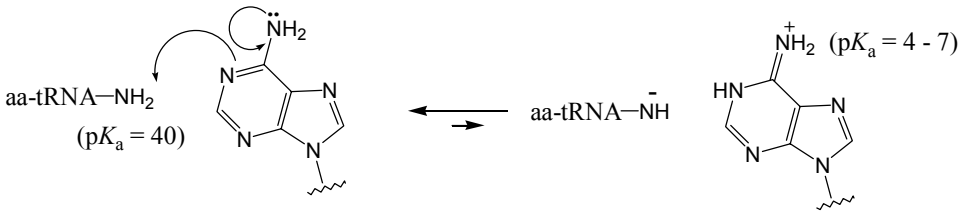
Proposed Mechanisms for Peptide Bond Formation

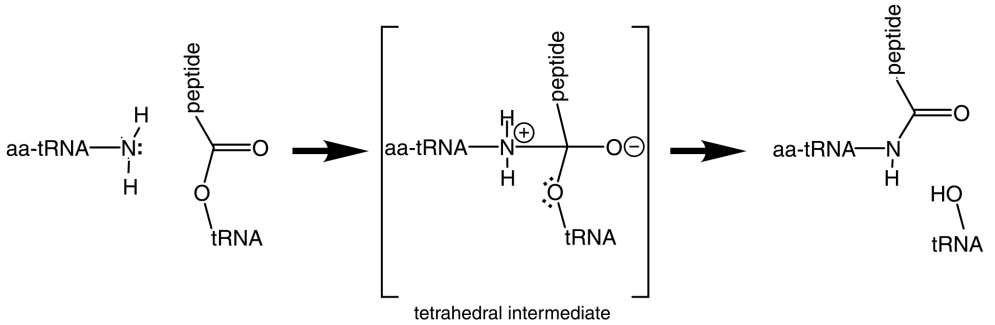
"During the early stages of model building on the 50S ribosome crystal structure, Moore, Steitz, and co-workers came to the anticipated, but hitherto unproven, conclusion that the peptidyl transferase center of the ribosome is composed exclusively of RNA." (see: *Science, August 2000, Vol. 289. no. 5481, pp. 905 - 920*).



from: Science, 2000 Vol. 289. no. 5481, pp. 947 – 950.

Problem:



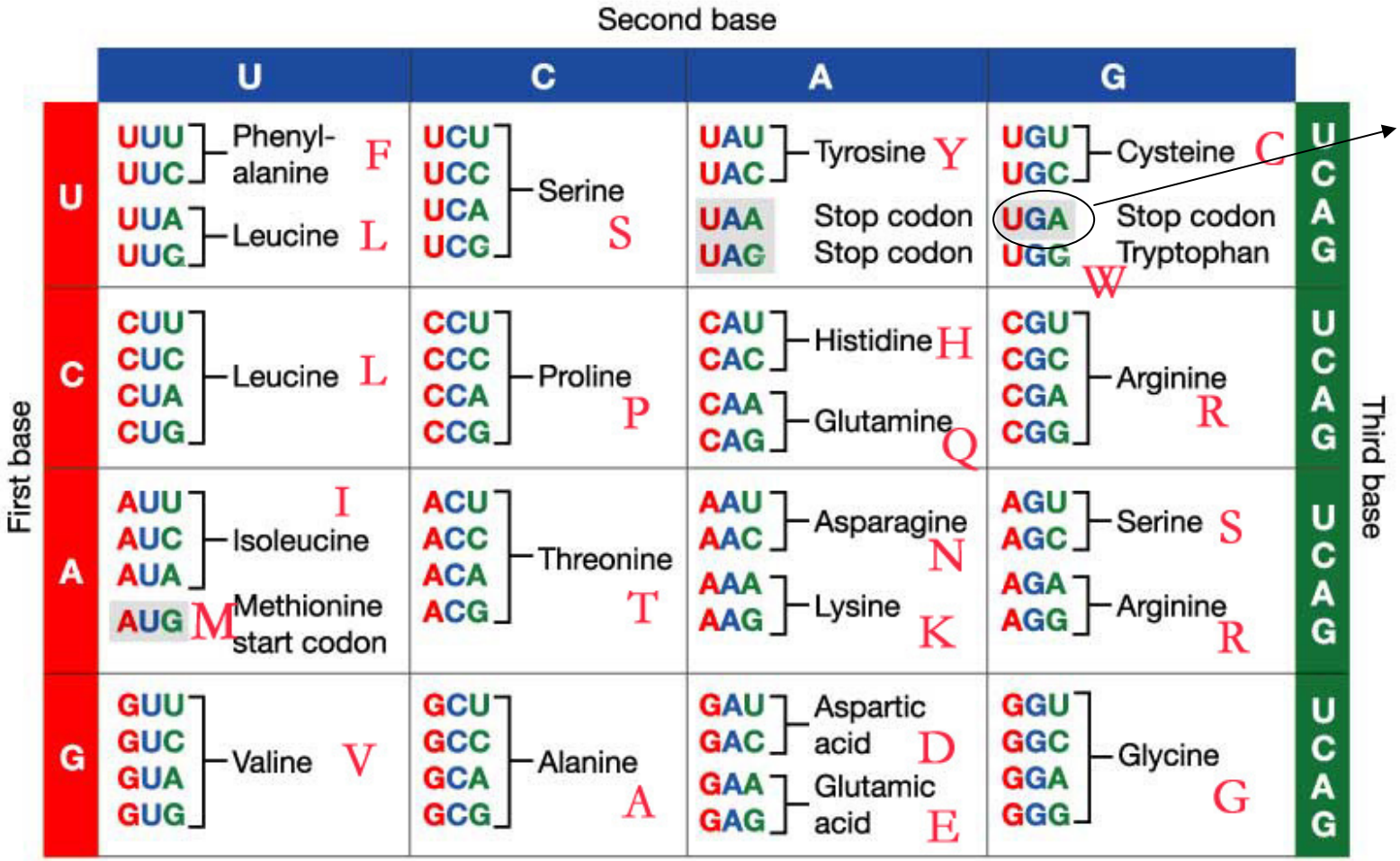


Revised mechanism from author's current website.

"…..We concluded that the proposed unusual pK_a of A2451, if existing, may not be crucial for the ribosome activity and that the previously reported pH-dependent alterations in the DMS modification of A2451 do not necessarily reveal an unusual pK_a of this nucleotide."

-Peter Sandler and Erik C. Bottger, et. al.
RNA, 2001, 7, 1365–1369.

Codon Table



Or, UGA can encode for a 21st amino acid like Sec, or an unnatural one.

Importance of third base?

Notice: three different stop codons.

http://teachline.ls.huji.ac.il/72693/codon.jpg