

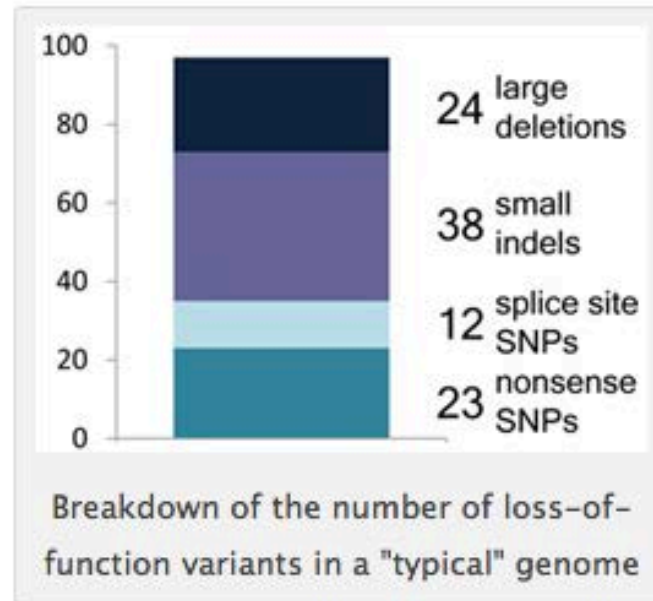
All genomes are dysfunctional: broken genes in healthy individuals

16/02/2012

Categories: [Journal Club](#)

Written by [Daniel MacArthur](#)

I don't normally blog here about my own research, but I'm making an exception for [this paper](#). There are a few reasons to single this paper out: firstly, it's in *Science* (!); and secondly, no fewer than five Genomes Unzipped members (me, Luke, Joe, Don and Jeff) are co-authors. For me it also represents the culmination of a fantastic postdoc position at the [Wellcome Trust Sanger Institute](#) (for those who haven't heard on Twitter, I'll be starting up a new research group at Massachusetts General Hospital in Boston next month).



Readers who don't have a *Science* subscription can access a pre-formatted version of the manuscript [here](#). In this post I wanted to give a brief overview of the study and then highlight what I see as some of the interesting messages that emerged from it.



***DNA: the ultimate
hard drive***

Digital Memories, Piling Up, May Prove Fleeting

By KATIE HAFNER (NYT) 1712 words

Published: November 10, 2004

The nation's 115 million home computers are brimming over with personal treasures -- millions of photographs, music of every genre, college papers, the great American novel and, of course, mountains of e-mail messages.

Yet no one has figured out how to preserve these electronic materials for the next decade, much less for the ages. Like junk e-mail, the problem of digital archiving, which seems straightforward, confounds even the experts.

"To save a digital file for, let's say, a hundred years is going to take a lot of work," said Peter Hite, president of Media Management Services, a consulting firm in Houston. "Whereas to take a traditional photograph and just put it in a shoe box doesn't take any work." Already, half of all photographs are taken by digital cameras, with most of the shots never leaving a personal computer's hard drive.

So dire and complex is the challenge of digital preservation in general that the Library of Congress has spent the last several years forming committees and issuing reports on the state of the nation's preparedness for digital preservation.

Jim Gallagher, director for information technology services at the Library of Congress, said the library, faced with "a deluge of digital information," had embarked on a multiyear, multimillion-dollar project, with an eye toward creating uniform standards for preserving digital material so that it can be read in the future regardless of the hardware or software being used. The assumption is that machines and software formats in use now will become obsolete sooner rather than later.

"It is a global problem for the biggest governments and the biggest corporations all the way down to individuals," said Ken Thibodeau, director for the electronic records archives program at the National Archives and Records Administration.

In the meantime, individual PC owners struggle in private. Desk drawers and den closets are filled with obsolete computers, stacks of Zip disks and 3 1/2-inch diskettes, even the larger 5 1/4-inch floppy disks from the 1980's. Short of a clear solution, experts recommend that people copy their materials, which were once on vinyl, film and paper, to CD's and other backup formats.

DIGITAL HISTORY
A GUIDE TO GATHERING, PRESERVING, AND PRESENTING THE PAST ON THE WEB
DANIEL J. COHEN AND ROY ROSENZWEIG

Home
Introduction
Exploring the History Web
Getting Started
Becoming Digital
Designing for the History Web
Building an Audience
Collecting History Online
Owning the Past?

Preserving Digital History
Introduction
The Fragility of Digital Materials

PRESERVING DIGITAL HISTORY

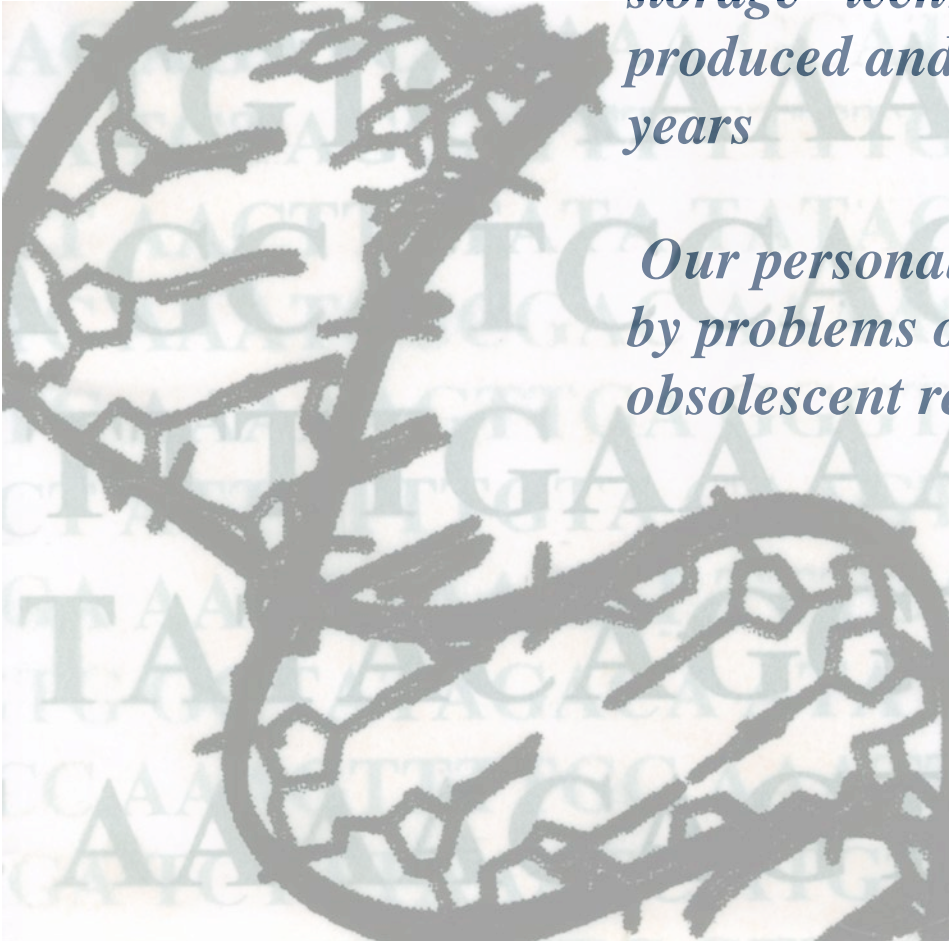
The Fragility of Digital Materials

If only digital preservation were as easy as changing the quality of the paper we print on, as publishers and archivists have done by using high-grade acid-free paper for documents deemed sufficiently important for long-term preservation. Electronic resources are profoundly unstable, far more unstable than such paper records. On the simplest level, many of us have experienced the loss of a floppy's or hard drive's worth of scholarship. The foremost American authority on the longevity of various media, the National Institute of Standards and Technology (NIST), still cannot give a precise timeline for the deterioration of many of the formats we currently rely on to store precious digital records. A recent report by NIST researcher Fred R. Byers notes that estimates vary from 30 to 300 years for popular media such as the CD and DVD, and

Data currently being stored in magnetic or optical media will probably become unrecoverable within a century or less (estimates vary). This will be due to the combined effects of

- 1. software obsolescence*
- 2. obsolescence of hardware for retrieval*
- 3. decay of the storage medium (aka material deterioration).*

This human-generated digital technology contrasts with the impressive information storage “technology” that evolution has produced and refined over the past 3+ billion years



Our personal digital archive is NOT plagued by problems of chemical fragility or by obsolescent retrieval systems

It can store the information from a million CDs in a space no bigger than your little finger, and could keep it safe for centuries. NPR 1/23/13

DNA has three properties that recommend it as a vehicle for long-term information storage:

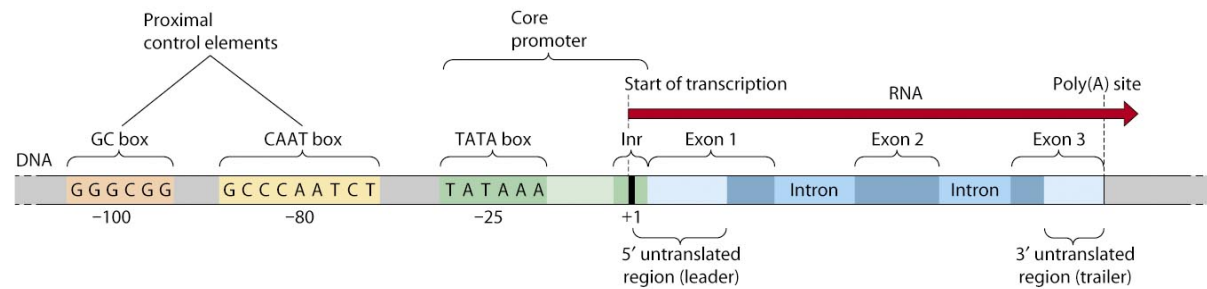
- **First**, *DNA has stood the informational "test of time" during the billions of years since life emerged. Non-replicating DNA molecules are quite robust.*
→ good chemical stability
- **Second**, because DNA is our genetic material, *methods for both storage and reading of DNA-encoded information is central to technological civilizations and will undergo continual improvements.*
→ DNA-R-US
- **Third**, use of DNA as a storage medium *permits each segment of information to be stored in an enormous number of identical molecules. This extensive informational redundancy would strongly mitigate effects of any losses due to stochastic decay.*
→ easy to make copies via PCR
- Data retrieval of information stored in DNA should ideally require minimal prior knowledge beyond a familiarity with molecular biological techniques (DNA sequencing and PCR --polymerase chain reaction)

The changing definition of the gene:

Mendelian: the fundamental functional unit of heredity that carries information from one generation to the next

Biochemical: a unit of heredity that specifies the production of a polypeptide

Molecular: a segment of DNA composed of a transcribed region and adjacent regulatory regions that control transcription



*permanent
archive of
genetic instructions*

DNA

→→

*short-term,
throw-away copy*

mRNA

→→

PROTEINS

Short stretches of
DNA are copied or
transcribed

**rRNA
tRNA
siRNA
miRNA**

Conversion of
genetic information
into a different
chemical form:
Translation from one
chemical language to
another

*Control of Biological
processes/Specification of
organism*

Our focus with respect to molecular genetics will be on:

- *how mutations come about*
- *how mutations affect gene function*
- *techniques to directly assess genotype at the DNA level*

***Sex, Errors and the Genome* by Mark Ridley**
Natural History (6/2001)

“At conception, human embryos average about 200 copying errors and about 50% of the embryos have a botched number of chromosomes. “

WHO IS TO BLAME?

When a thirty year old man breeds
with a 30 year old woman:

- his DNA (in his sperm cells) has been copied *430 times* against her *33 cell division* (in egg cells).
- with thirteen times as many errata in his DNA, about 185 of the 200 copying mistakes in each human conception may come from the sperm.
- however, a woman's eggs are more likely to carry serious errors in chromosome numbers, and these errors increase with maternal age.

Sex, Errors and the Genome by Mark Ridley Natural History (6/2001)

MUTATIONS: MOTHER VERSUS FATHER

As their life spans stretch out, men and women travel different evolutionary roads, and the amount of DNA copying that goes on in their gonads contributes to the error level of their genomes in different ways. Men manufacture sperm throughout their lives. About 40 cell divisions in the reproductive cells have occurred in a human male by the time he reaches puberty. After that, the DNA in his sperm is copied every sixteen days, or 23 times per year. A twenty-year-old man's genome has been copied more than 200 times, and a forty-year-old's more than 600 times. Compare that with the average adult male rat: its DNA has been copied only 58 times in its short life, and the DNA in its spermatozoa is therefore relatively error free.

A female human, on the other hand, already possesses her lifetime supply of eggs--with about 33 cell divisions behind them--by the time she is a late-stage fetus. When a thirty-year-old man breeds with a thirty-year-old woman, his DNA has been copied 430 times against her 33. With about thirteen times as many errata in his DNA, about 185 of the 200 copying mistakes in each human conception may come from the sperm. However, a woman's eggs are more likely to carry serious errors in chromosome numbers, and these errors increase with maternal age. Some disorders, such as Down syndrome, are the result of eggs that deliver the wrong number of chromosomes during conception.

All the DNA messages in a sperm and an egg can be compared with all the text in two sets of encyclopedias. If publishers made errors in book production at the same rate fathers and mothers do in transcribing their DNA, buyers of Britannica would receive sets with 200 printing errors on average, and half the time they'd be sent the wrong number of books.

MUTATION JARGON

GENE MUTATION = POINT MUTATION

(scale of mutation is small and is localized to a specific region,
a single nucleotide or a few adjacent base pairs)

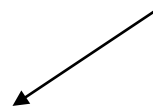
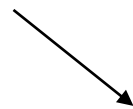


at the DNA level:

- ✧ single base pair substitutions: *transitions & transversions*
- ✧ single (or a few) base pair addition or deletion: *indels*
- ✧ gene mutation by transposon insertion

at the level of
gene expression:
promoter mutations
splicing mutations
regulatory mutations

at the protein
level:
nonsense
missense
[neutral]
silent
frameshift

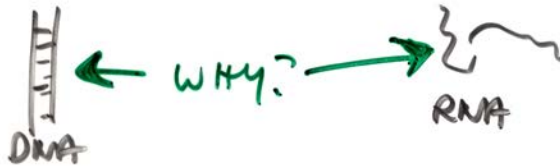
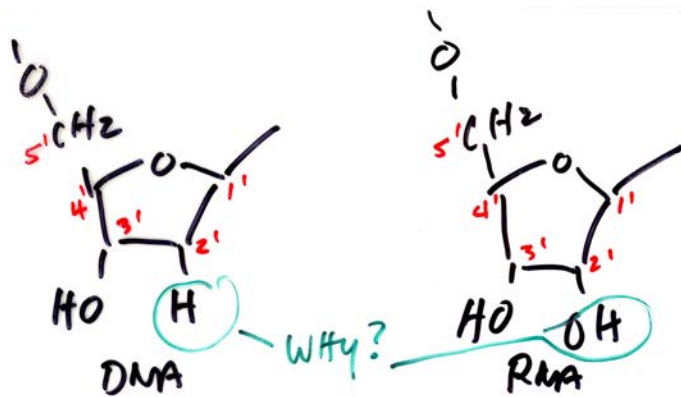


at the level of gene function:

loss-of-function
gain-of-function
[neutral]

CHROMOSOME MUTATION

- involves *segments of chromosomes* or *whole chromosomes* or *whole genomes*
- alterations in chromosome structure and number
- deletion, duplications, translocations and inversions
- CNVs: copy number variations



PYRIMIDINE
BASES

PYRIMIDINE
BASES

THYMININE

CYTOSINE

DNA

WHY?

URACIL

CYTOSINE

RNA

WHICH CAME
FIRST?

Review DNA structure and DNA replication

- general overview
- biochemistry of chain elongation
- features of DNA polymerases

5 ' TTACCCATTCAGCCCATTCCCTGCAAACCAGTGGAGTATCCGCTGCAGCTGCTGCACAGCCCCCTGCCCCAGTGGTGAAGAGGCC
TGGGGCCATGGCCACCCACCACCCCCTGCAGGAGCCCTCCCAGCCCCTGAACCTCACAGCCAAGCCCAAGGCCCCCGAGCTGCCCAACA
CCTCCAGCTCCCCAAGCCTGAAGATGAGCAGCTGTGTGCCCCGCCCCCCCAGCCATGGAGGCCCCACGCGGGACCTGCAGTCCAGCCCC
CCGAGCCTGCCTCTGGGCTTCCTTGGTGAAGGGGACGCTGTCACCAAAGCCATCCAGGATGCTCGGCAGCTGC..... .
..... etc, etc, etc, 3 '

The coding information contained in DNA is

- **one dimensional:** encoded along the length of a molecule
- **digital:** because the basic information unit (the nucleotide or base) can exist in only one of four* discrete states abbreviated: G, A, T, C

Review DNA structure here:

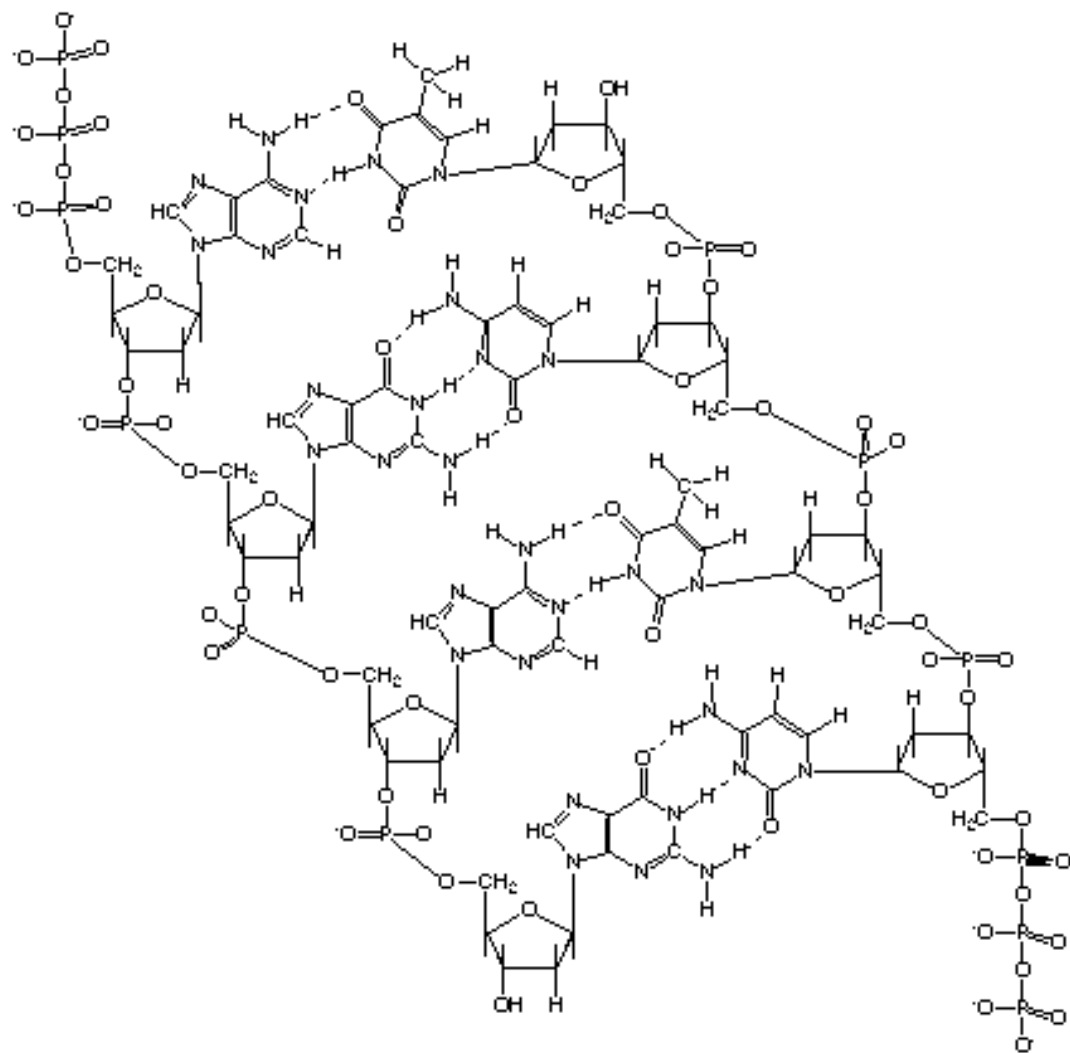
<http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/D/DoubleHelix.html>

Cool stuff here

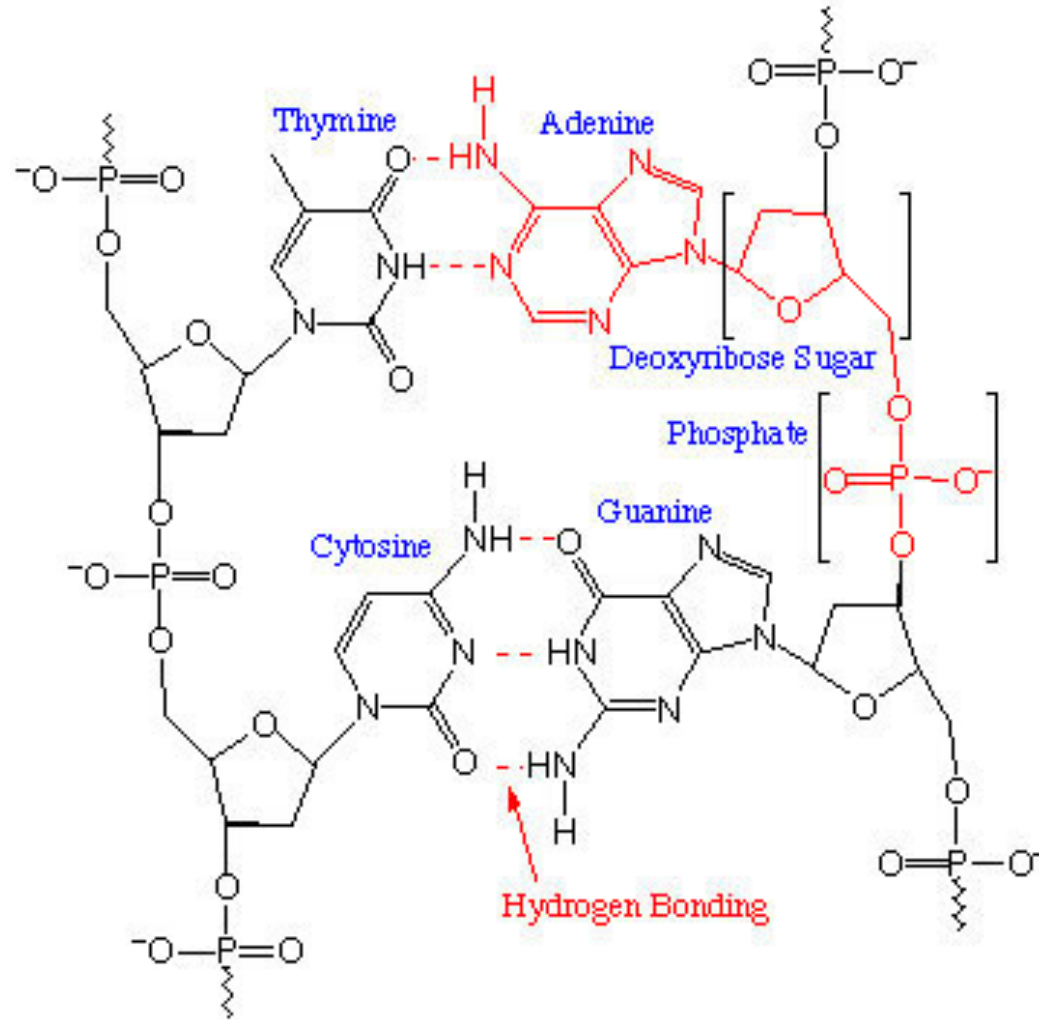
<http://www.johnkyrk.com/DNAanatomy.html>

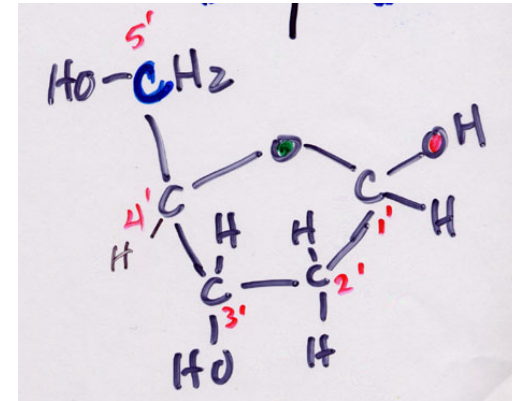
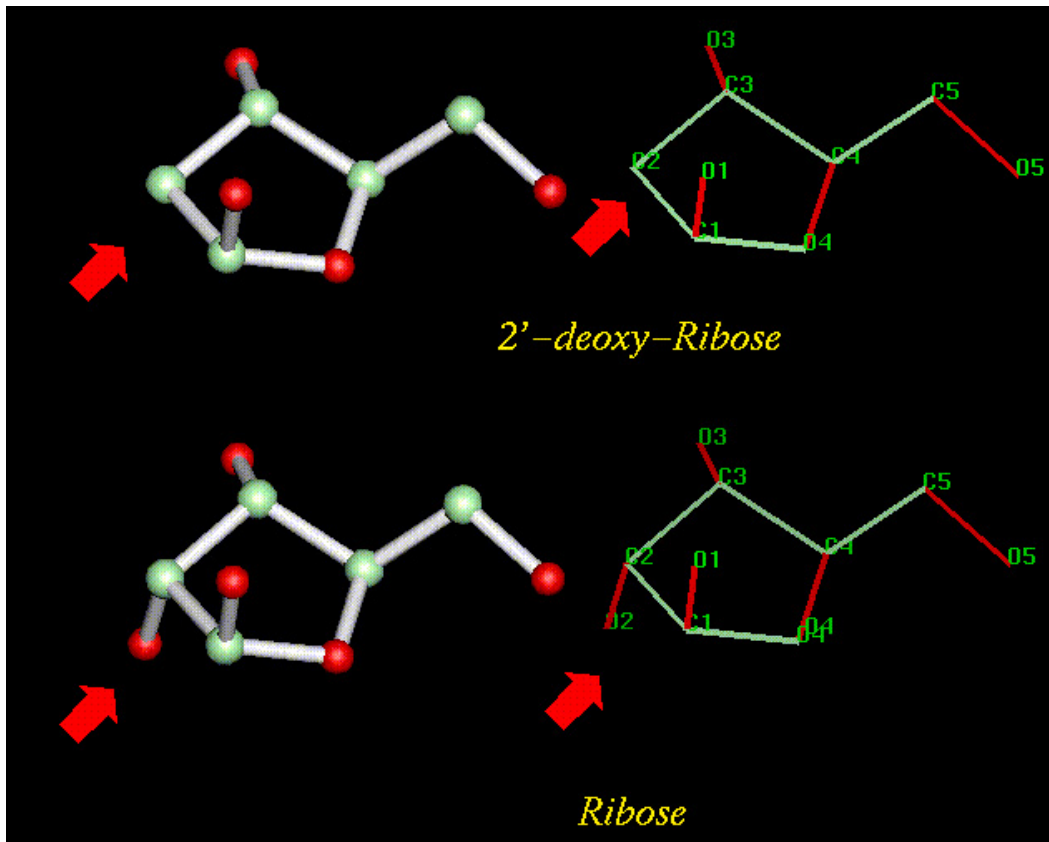
<http://www.geneticengineering.org/chemis/Chemis-NucleicAcid/DNA.htm>

***DUH -- somebody has realized that more info could stored on a CD if each point (position) could be represented by more than two (0 or 1) possibilities .**



2-D look at DNA: note strands are *antiparallel*

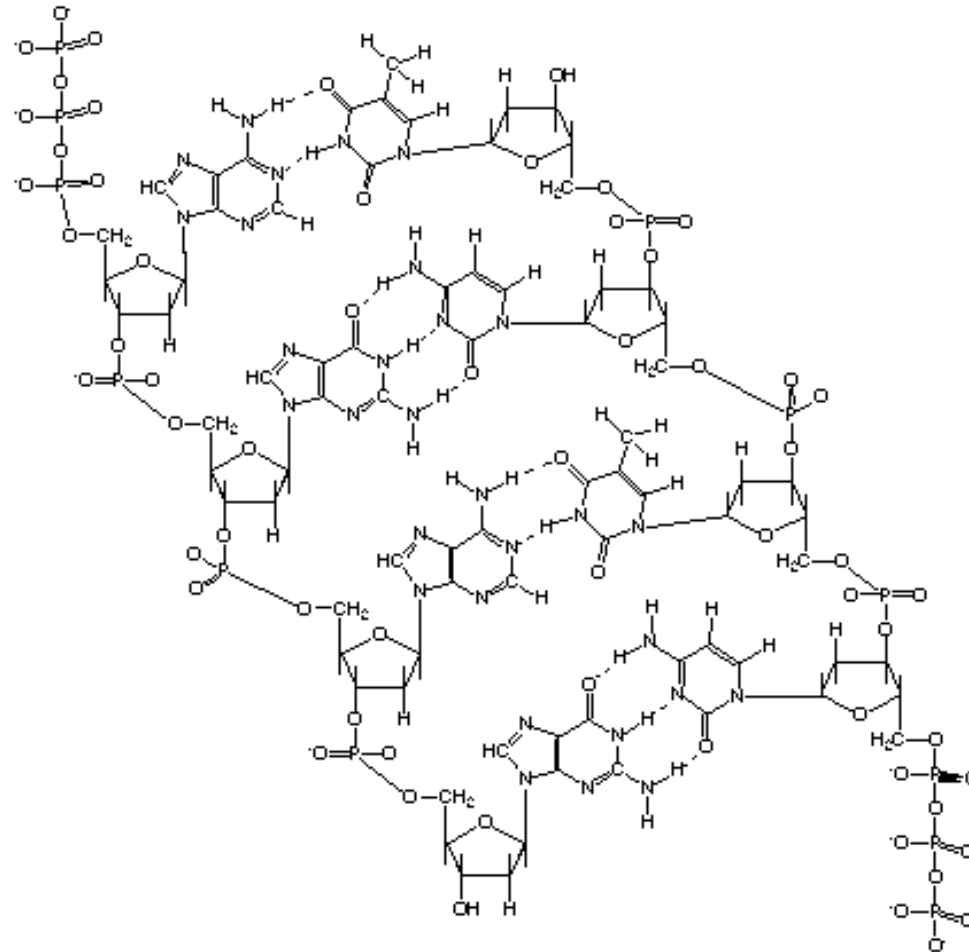


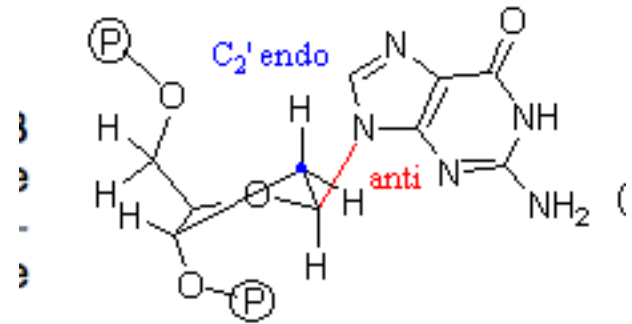
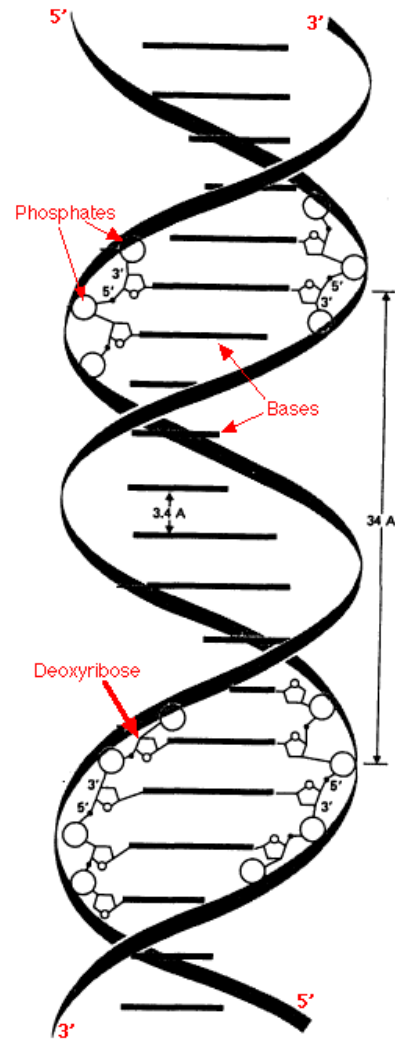
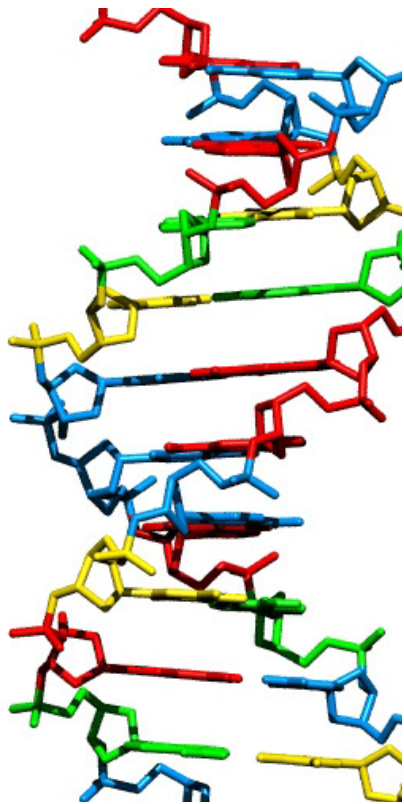


WHY 2' deoxy in DNA?

THE ENDS OF A DNA POLYMER CAN BE DISTINGUISHED BASED ON WHETHER THERE IS A FREE 5' OR 3' HYDROXYL

- 3' and 5' ends
- antiparallel



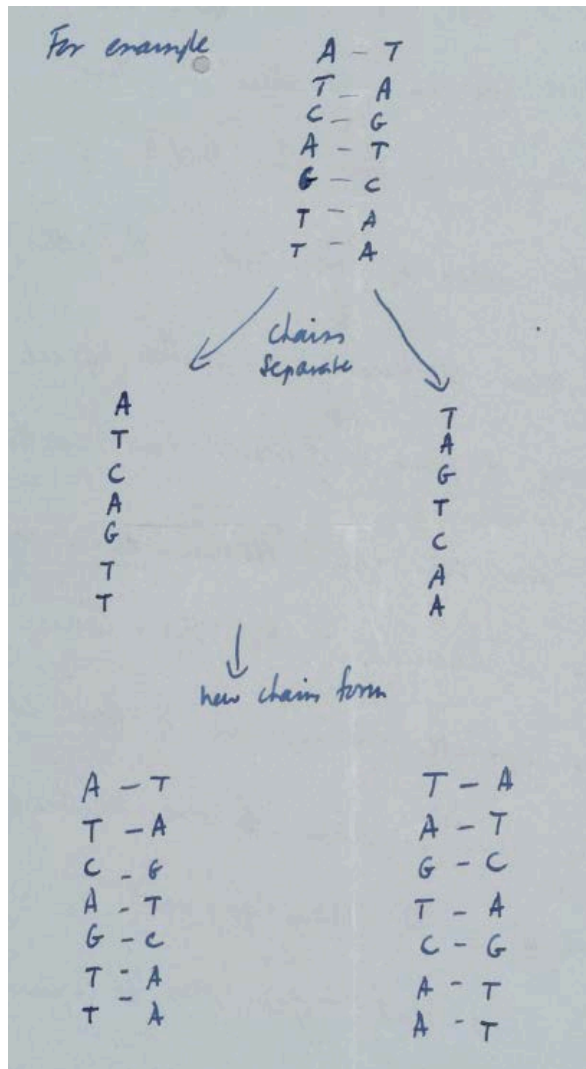




What is Crick holding in his hand?

<http://biocrs.biomed.brown.edu/Books/Chapters/Ch%208/DH-Paper.html>

<http://www.nature.com/genomics/human/watson-crick/>



Semi-conservative DNA replication:

1. The parental strands of the DNA double helix separate
2. Each parental strand serves as template for the synthesis of a complementary copy
3. The nucleotide sequence of the newly synthesized daughter strand is determined by the
 - sequence of the parental template
 - pairing (hydrogen-bonding) specificities of the purine and pyrimidine bases

← *Francis Crick explains his idea about DNA replication in a letter to his son Michael dated 19 March 57. See letter at this link:*

<http://fire.biol.wvu.edu/trent/trent/CricktoMichael.pdf>

Nice animation:

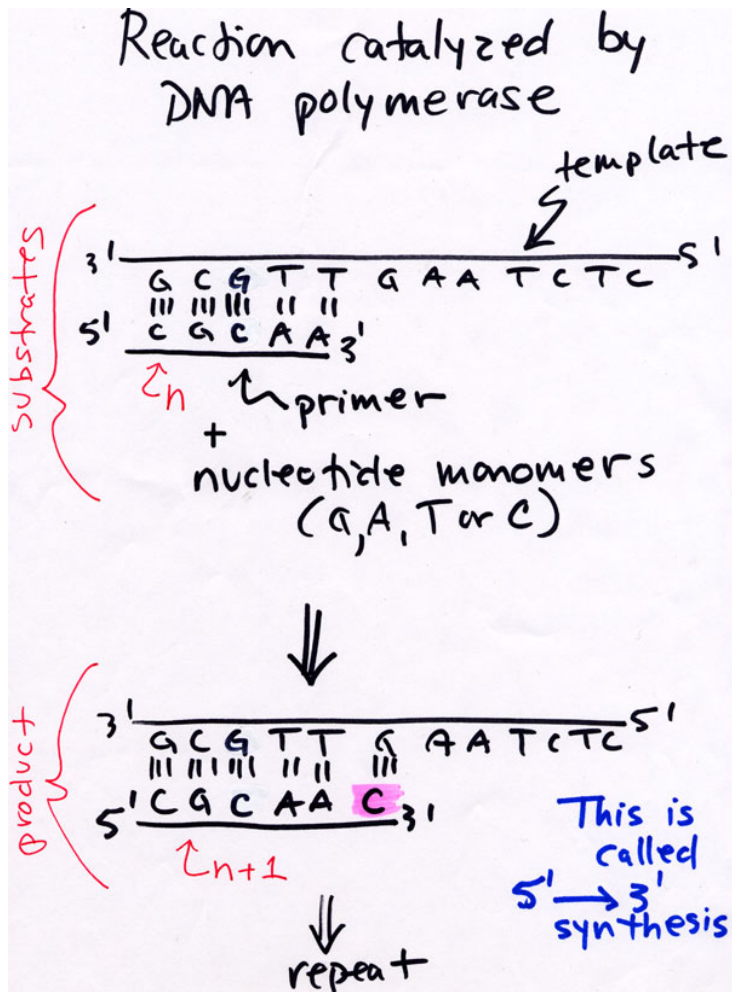
<http://www.johnkyrk.com/DNAreplication.html>

What do you know about the enzymology of this process?

What is the name of the enzyme that catalyzes the elongation of the nucleotide chain?

***DNA polymerase*: synthesizes new strands of DNA by catalyzing the chain elongation reaction: one nucleotide is added to the existing DNA chain with the release of inorganic phosphate**

- **Synthesis of the polymer is always 5' ---> 3'**
- **A template is required**
- **A primer is required**



<http://www.geneticengineering.org/chemis/Chemis-NucleicAcid/DNA.htm>

Nice animation:

<http://www.johnkyrk.com/DNAreplication.html>

Figure shows features common to all DNA polymerases:

- it takes “instructions” from a template -- the parental strand of DNA
- it can only catalyze the addition of a nucleotide monomer to a 3’ carbon of ribose
- it cannot catalyze the addition of a nucleotide monomer to the 5’ carbon of ribose
- this is called 5’ to 3’ synthesis

DNA polymerase cannot lay down the first nucleotide of a DNA strand

- It requires a “bit” of polymer to add onto
- This short segment of polymer is called a *primer*
- It provides a 3' hydroxyl for the DNA polymerase to add onto

During DNA replication *in the cell (in vivo)*:

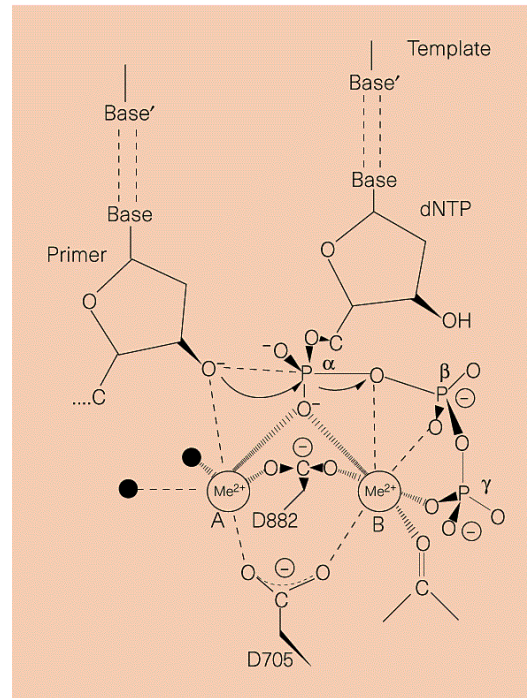
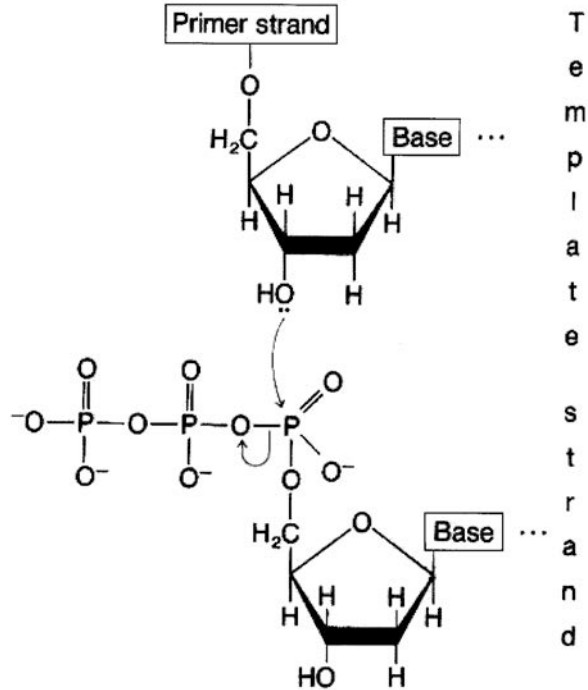
- new DNA chains are initiated with an RNA primer to which the newly synthesized DNA is attached
- then, the growing DNA chain acts as the primer

DNA synthesis *in the test tube (in vitro)*:

- primers are short polymers of DNA (oligonucleotides)

What is the significance of the primer requirement?

What happens if there is no primer?



DNA synthesis occurs in the 5' to 3' direction

The monomer substrates are in the form of a dNTP
d = 2' deoxy N = A, C, G or T TP = triphosphate

dNTP's are chemically reactive monomer units

Note again that it is the 3' end of the primer chain that forms the bond

The Fidelity of DNA Replication

✦ *A species genome is the record of instructions specifying the assembly and functioning of the organism*

✦ *Propagation of the species requires the accurate copying over (replication) of this set of instructions*

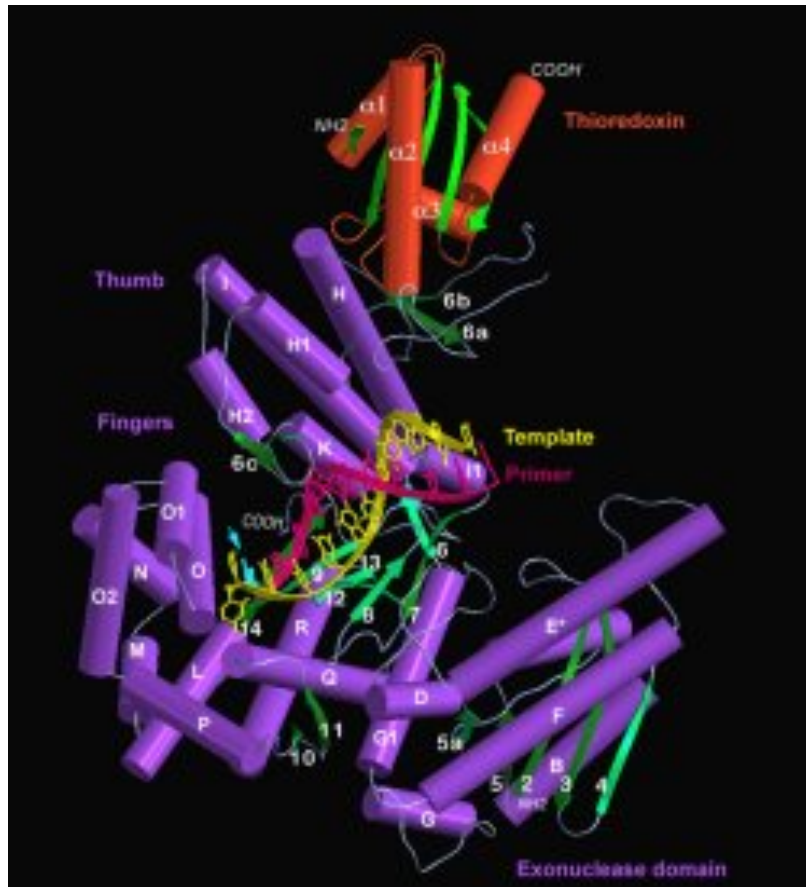
✦ *For organisms with large complex genomes, attaining sufficient accuracy is an impressive feat*

✦ *Studies have shown that high complexity of a genetic program necessitates a correspondingly high accuracy of copying for it to be transmitted to offspring as a meaningful program.*

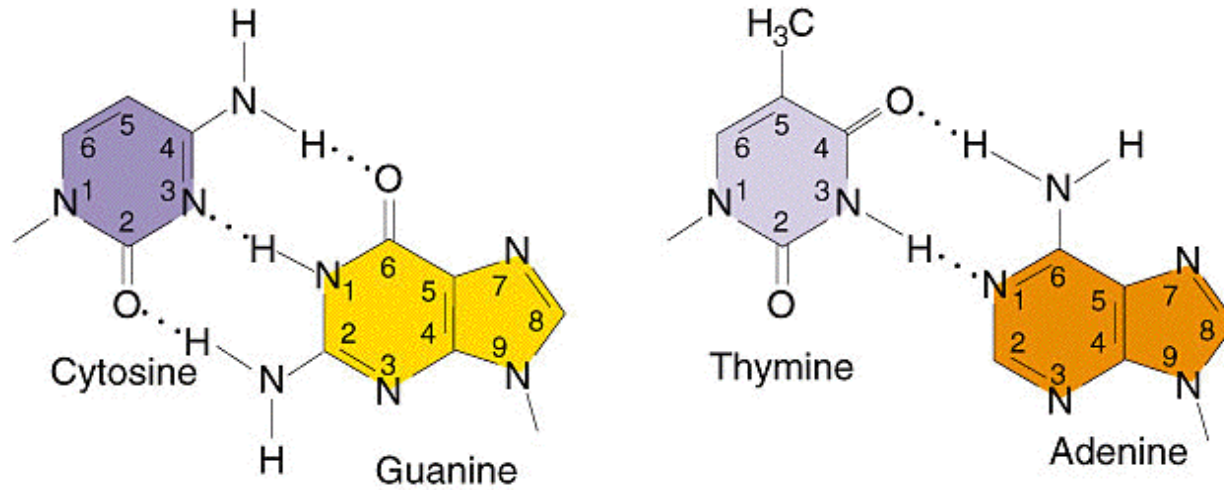
✦ *For a given genomic complexity, there is a critical-copying fidelity below which information cannot longer be maintained.*

Cell's strategy for getting a high-quality DNA replication product?

- *Nucleotide Selection: do a good job to begin with* (some inherent limitations of the base-pairing specificities though)
- *Proofreading: perform an immediate double-check* of your work as you go along
- *Post-replication mismatch repair: go back and double-check your work again* against the original copy after you've completed the task

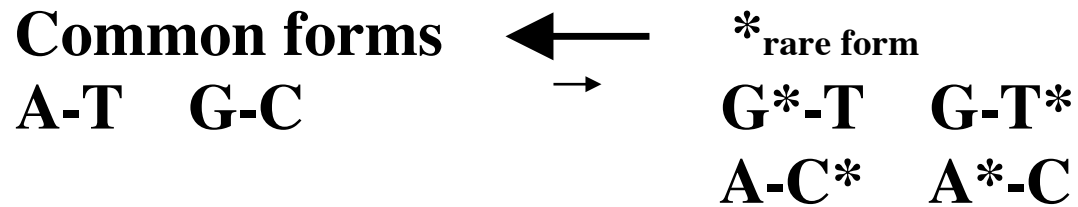


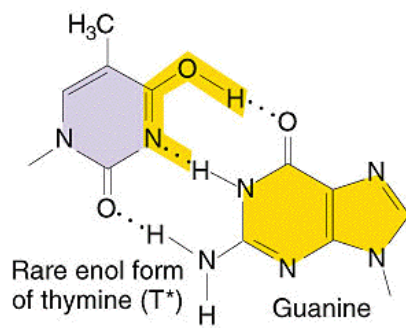
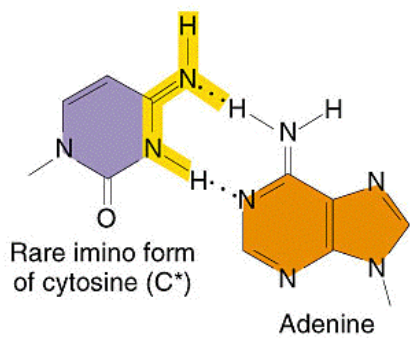
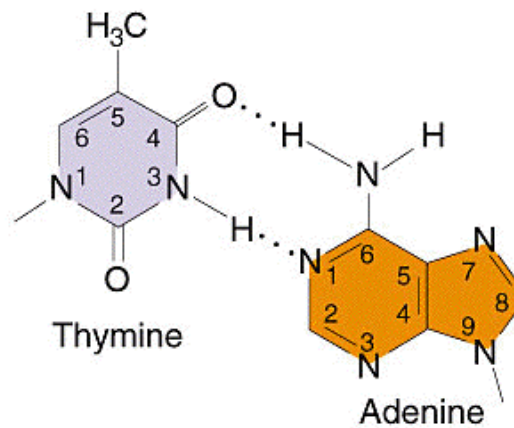
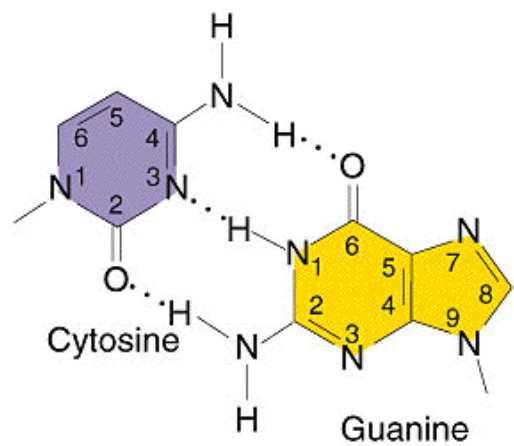
Quality control mechanisms:
(1) *Nucleotide selection*: correct pairing is the most energetically favorable; that is, the complex of polymerase, template DNA and nucleotide triphosphate is the most stable when the nucleotide bound is complementary to the template nucleotide.



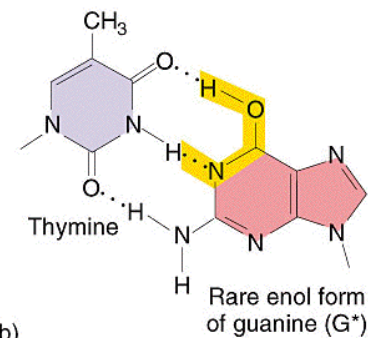
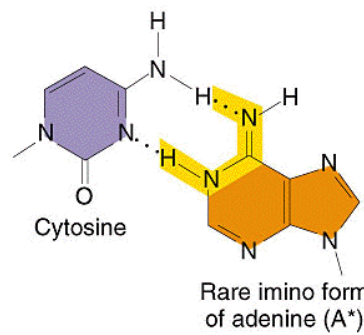
How specific is this interaction: GC and AT?
How many mistakes are made by DNA polymerase at this step?

In a DNA molecular inside the cell, the bases are in an equilibrium between a stable form (shown in previous figure) and an unstable form (next figure)





(a)



(b)

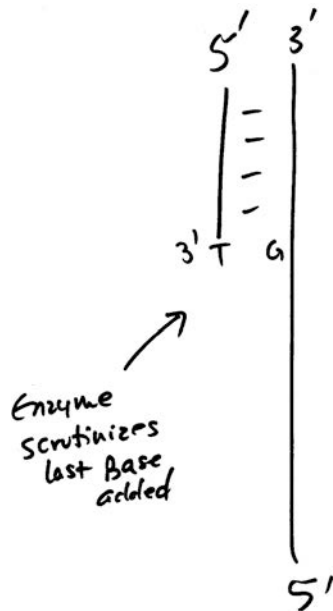
Mistake in nucleotide selection can occur when a base is in its rare form

Error rate at this initial polymerization step: *1 in 10^4 - 10^6 nucleotides incorporated is non-complementary*

The error rate at nucleotide selection seems pretty good. *Is it?*

(2) Proofreading by the DNA polymerase: the polymerase edits the DNA sequence by removing mispaired nucleotides that have been incorrectly inserted during the polymerization reaction.

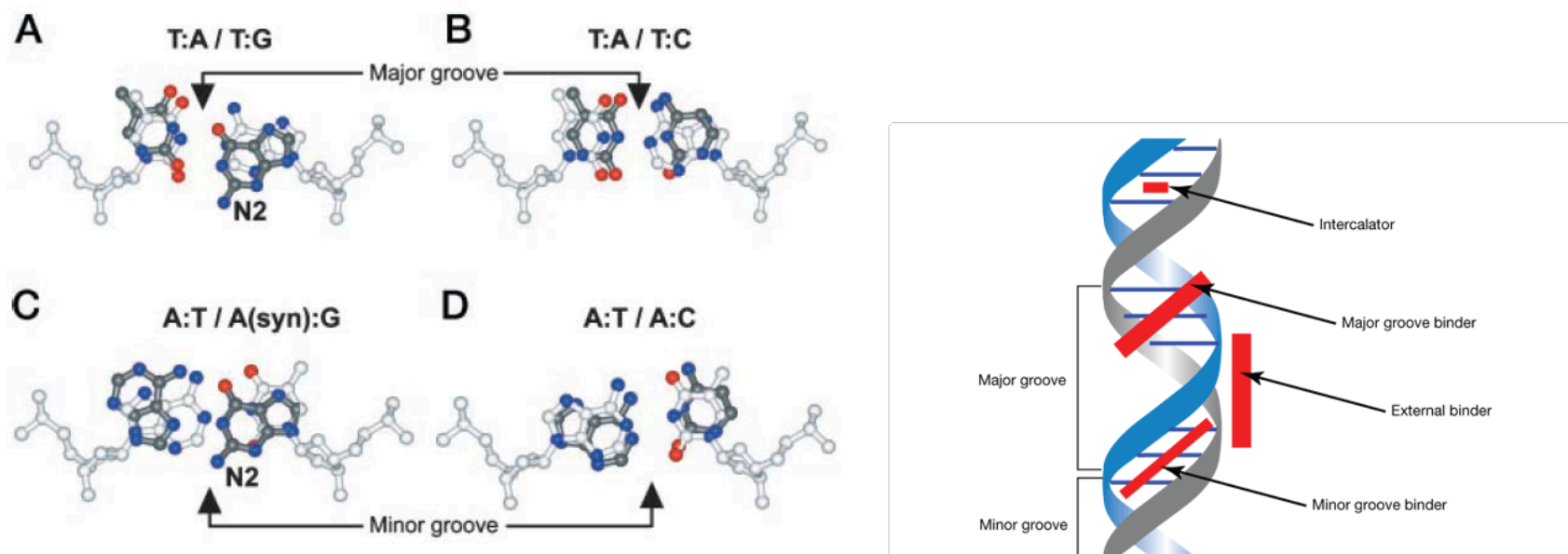
DNA polymerase has a 3'-5' exonuclease function that acts as a proofreader: the enzyme “scrutinizes” the most recently added nucleotide and removes it if it is non-complementary



Mismatched bases are detected as they have weaker bonding interactions—the 'melting' temperature is lower—and this increases the chance of switching from the polymerase to the exonuclease active site

When DNA polymerase stalls at a mismatched base, the proofreading function kicks in

Superpositions of the structures of Watson-Crick basepairs and mispairs *Base Substitution Specificity of DNA Polymerase β Depends on Interactions in the DNA Minor Groove*: unfavorable interactions between an active site arginine side chain and mispair-specific atoms in the minor groove contribute to DNA polymerase specificity.



THE JOURNAL OF BIOLOGICAL CHEMISTRY Vol. 274, pp. 20749–20752, 1999

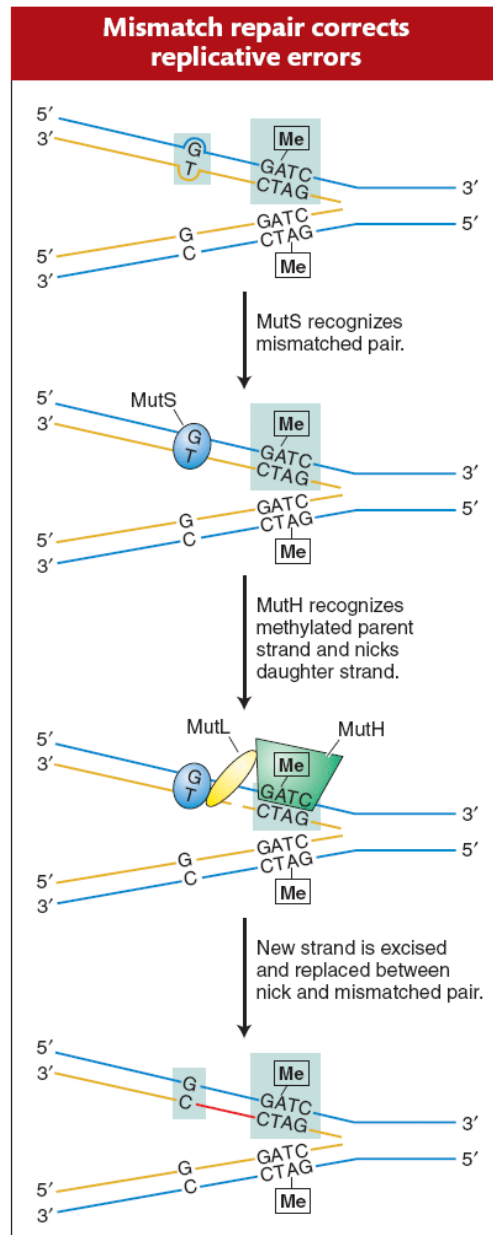
\ Each panel shows a ball-and-stick representation of a correct Watson-Crick base pair superimposed with the mispair indicated (as described under “Experimental Procedures”). The Watson-Crick base pairs are shown in white, and the bases of the mispaired nucleotides are shown in gray. The oxygen atoms of the bases are depicted as red balls and the nitrogen atoms as blue balls. The major and minor grooves of the DNA helix are indicated. A, comparison of a correct T-A base pair with a T-G wobble base pair; B, the T-A base pair compared with a T-C mispair; C, comparison of a Watson-Crick A-T base pair with the A(syn)-G pair; D, the A-T base pair compared with an A-C mispair.

(3) Post-synthesis correction mechanisms: correction of errors that remain after proofreading. One example is mismatch repair which correct mismatches in newly synthesized DNA by

- (a) detecting a mismatched base pair,
- (b) removing the mismatched nucleotide,
- (c) replacing it with the correct nucleotide.

How does the enzymatic machinery know which of the two mismatched bases to correct?

Figure 15-26 in 9th
Figure 16-23 in 10th



Post replication mismatch repair in *E. coli*

Mut = mutator

The components of mismatch repair systems are *very highly conserved from bacteria to man*

Eukaryotes use different mechanisms to distinguish between the parental and daughter strands (but as of 2007 were yet undefined)

Inherited mutations in the mismatch repair system predispose an individual to colon cancer (HNPCC)

Net (final) error rate after post-synthesis correction is estimated to be: 1 in 10^9 - 10^{10}

(one mistake per one billion to ten billion nucleotides replicated)

*Interestingly, the final error rate (# mistakes per nucleotide replicated) appears to be the same in *E. coli* and humans despite the very large difference in genome size*

➡ *Some geneticists suggest that natural selection may have exhausted most the general ways of maintaining genetic fidelity -- in other words, this is a good as it gets given the inherent noise in any biochemical process*

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach,^{1*} Gustavo Glusman,^{1*} Arian F. A. Smit,^{1*} Chad D. Huff,^{1,2*} Robert Hubley,¹
Paul T. Shannon,¹ Lee Rowen,¹ Krishna P. Pant,³ Nathan Goodman,¹ Michael Bamshad,⁴
Jay Shendure,⁵ Radoje Drmanac,³ Lynn B. Jorde,² Leroy Hood,^{1†} David J. Galas^{1†}

We analyzed the whole-genome sequences of a family of four, consisting of two siblings and their parents. Family-based sequencing allowed us to delineate recombination sites precisely, identify 70% of the sequencing errors (resulting in >99.999% accuracy), and identify very rare single-nucleotide polymorphisms. We also directly estimated a human intergeneration mutation rate of $\sim 1.1 \times 10^{-8}$ per position per haploid genome. Both offspring in this family have two recessive disorders: Miller syndrome, for which the gene was concurrently identified, and primary ciliary dyskinesia, for which causative genes have been previously identified. Family-based genome analysis enabled us to narrow the candidate genes for both of these Mendelian disorders to only four. Our results demonstrate the value of complete genome sequencing in families.

30 APRIL 2010 VOL 328 **SCIENCE** www.sciencemag.org

*Is this frequency inconsistent with
the error rates on the previous page?*

The above error rate is that observed/estimated for replication of genomic DNA in eukaryotes and prokaryotes.

- *In contrast the error rate during replication of many viral genomes is much higher, probably due to the fact that their replication enzymes generally don't have error correction capabilities.*
- For example, the error rate for replication of the HIV (AIDS virus) genome is 1×10^{-4} or one mistake in every 10,000 bases copied.
- This high error rate certainly contributes to the high mutation rate of this virus.

Optional Reading Assignment

For your personal enrichment, see this thoughtful essay on biological processes and fidelity Nature 413: 115 Sept. 13, 2001

<http://fire.biol.wvu.edu/trent/trent/fidelity.pdf>