

# **Gene Structure & Gene Finding: Part I**

**David Wishart**

**Rm. 3-41 Athabasca Hall**

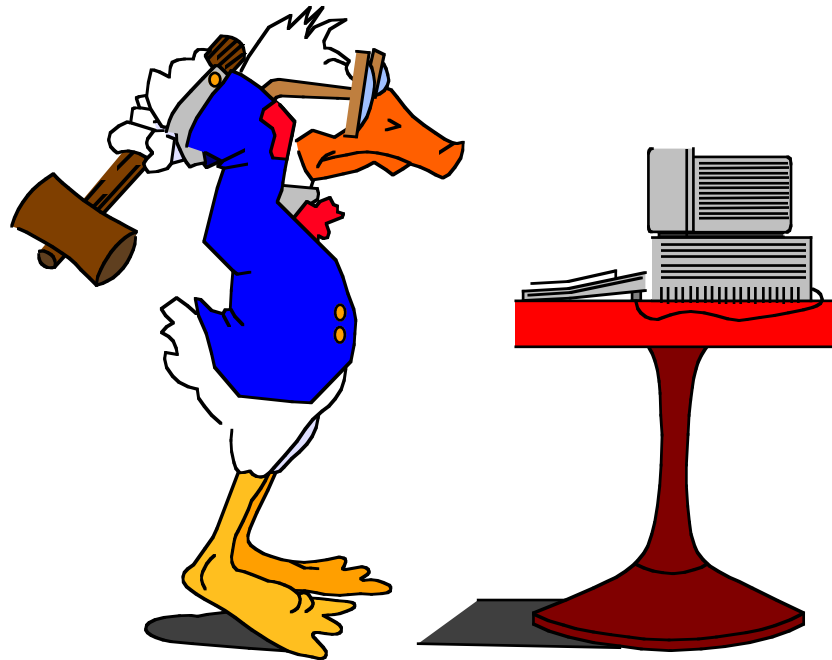
**[david.wishart@ualberta.ca](mailto:david.wishart@ualberta.ca)**

# Contacting Me...

- **200 emails a day – not the best way to get an instant response**
- ***Subject line: Bioinf 301 or Bioinf 501***
- **Preferred method...**
  - **Talk to me after class**
  - **Talk to me before class**
  - **Ask questions in class**
  - **Visit my office after 4 pm (Mon. – Fri.)**
  - **Contact my bioinformatics assistant – Dr. An Chi Guo ([anchigu@gmail.com](mailto:anchigu@gmail.com))**

# Lecture Notes Available At:

- <http://www.wishartlab.com/>
- *Go to the menu at the top of the page, look under Courses*



# Outline for Next 3 Weeks

- **Genes and Gene Finding (Prokaryotes)**
- **Genes and Gene Finding (Eukaryotes)**
- **Genome and Proteome Annotation**
- **Fundamentals of Transcript Measurement**
- **Introduction to Microarrays**
- **More details on Microarrays**

# My Lecturing Style

- **Lots of slides with limited text (room to add notes to the slides based on verbal information)**
- **If you don't show up to the lectures you'll miss most of the verbal information (sure to fail)**
- **Bioinformatics is mostly done on the web, key is knowing where to go and how to use websites**
- **I want you to spend some time (15-20 min) after each lecture to try/test the websites on your own**
- **Assignments build on what you've learned in class but also are intended to make you learn additional material to greater depth**

# Assignment Schedule

- **Gene finding - genome annotation**
  - (Assigned Oct. 31, due Nov. 7)
- **Microarray analysis**
  - (Assigned Nov. 7, due Nov. 19)
- **Protein structure analysis**
  - (Assigned Nov. 21, due Nov. 28)

**Each assignment is worth 5% of total grade, 10% off for each day late**

# Objectives\*

- **Review DNA structure, DNA sequence specifics and the fundamental paradigm**
- **Learn key features of prokaryotic gene structure and ORF finding**
- **Learn/memorize a few key prokaryotic gene signature sequences**
- **Learn about PSSMs and HMMs**
- **Learn about web tools for prokaryotic gene identification**

**Slides with a \* are ones that are important (could be on the test)**

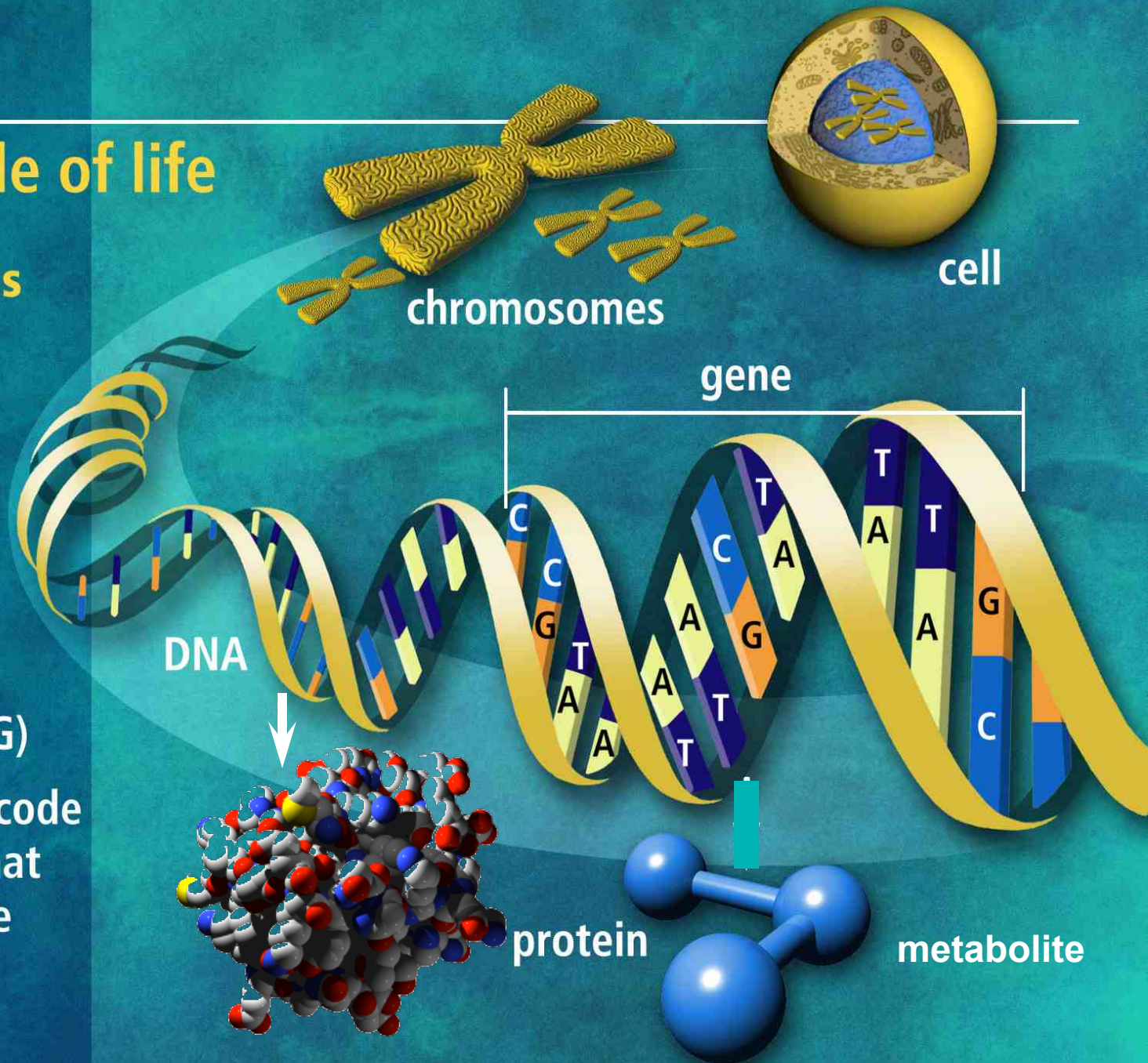
# DNA

## the molecule of life

### Trillions of cells

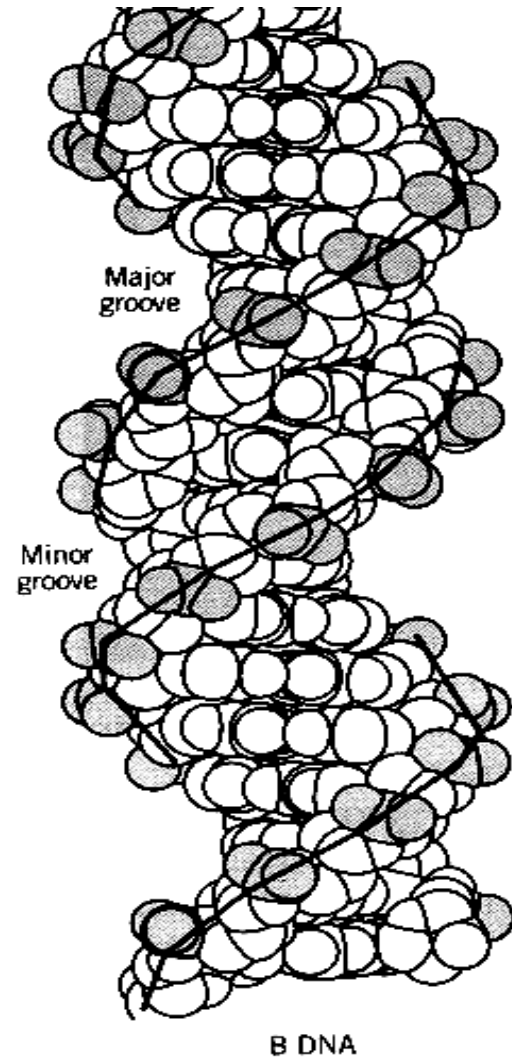
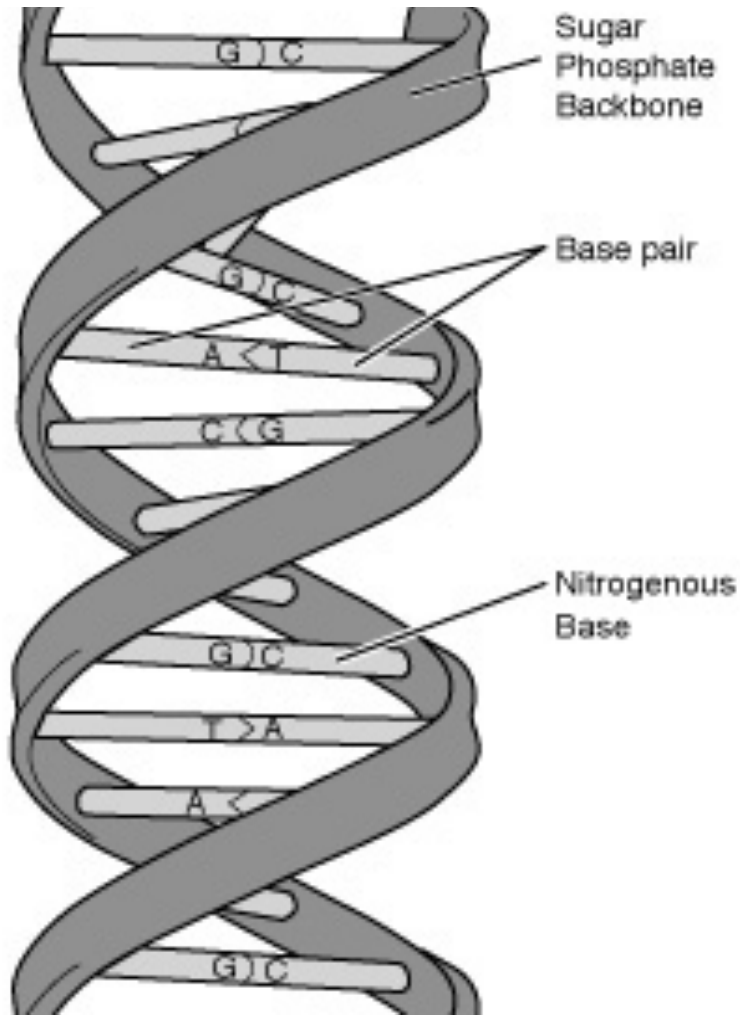
#### Each cell:

- 46 human chromosomes
- 2 m of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- **23,000** genes code for proteins that perform all life functions



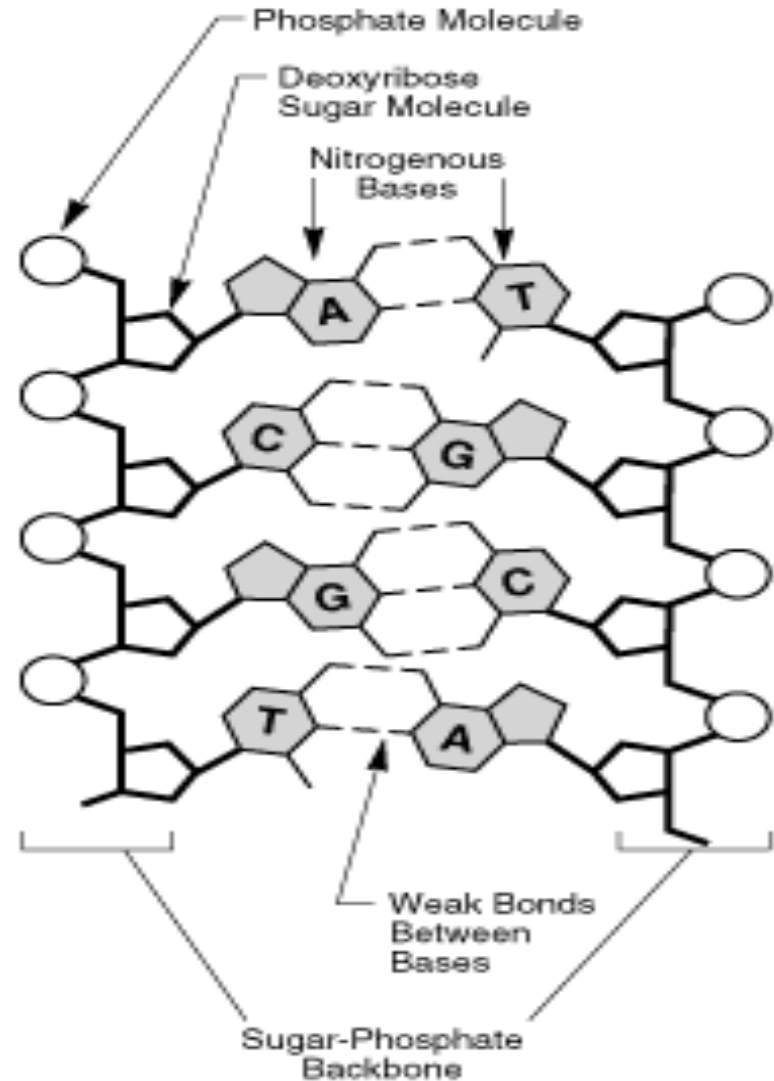


# DNA Structure



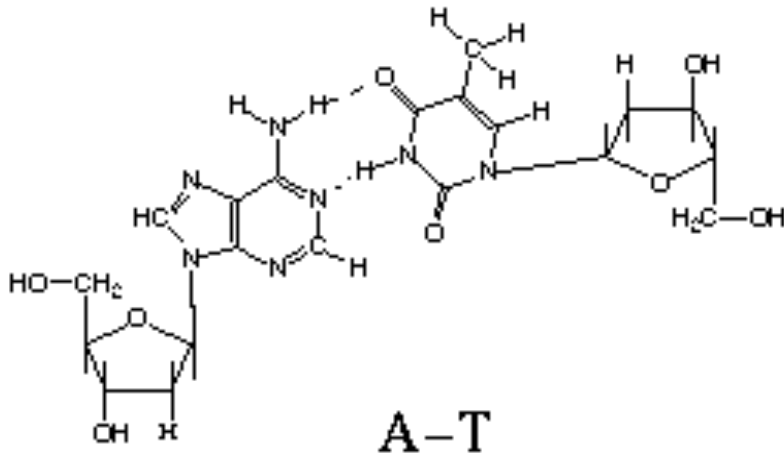
# DNA - base pairing\*

- **Hydrogen Bonds**
- **Base Stacking**
- **Hydrophobic Effect**



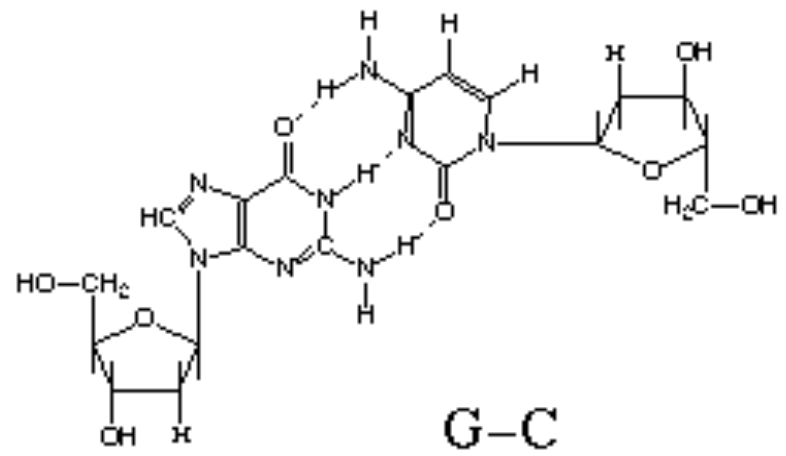
# Base-pairing (Details)\*

## DNA Basepairs



Adenosine-Thymidine  
(Adenine-Thymine)

**2 H-bonds**



Guanosine-Cytidine  
(Guanine-Cytosine)

**3 H-bonds**

# DNA Sequences

5'

3'

**Single:** ATGCTATCTGTACTATATGATCTA

5'

3'

**Paired:** ATGCTATCTGTACTATATGATCTA  
TACGATAGACATGATATACTAGAT

Read this way----->

5'

3'

ATGATCGATAGACTGATCGATCGATCGATTAGATCC

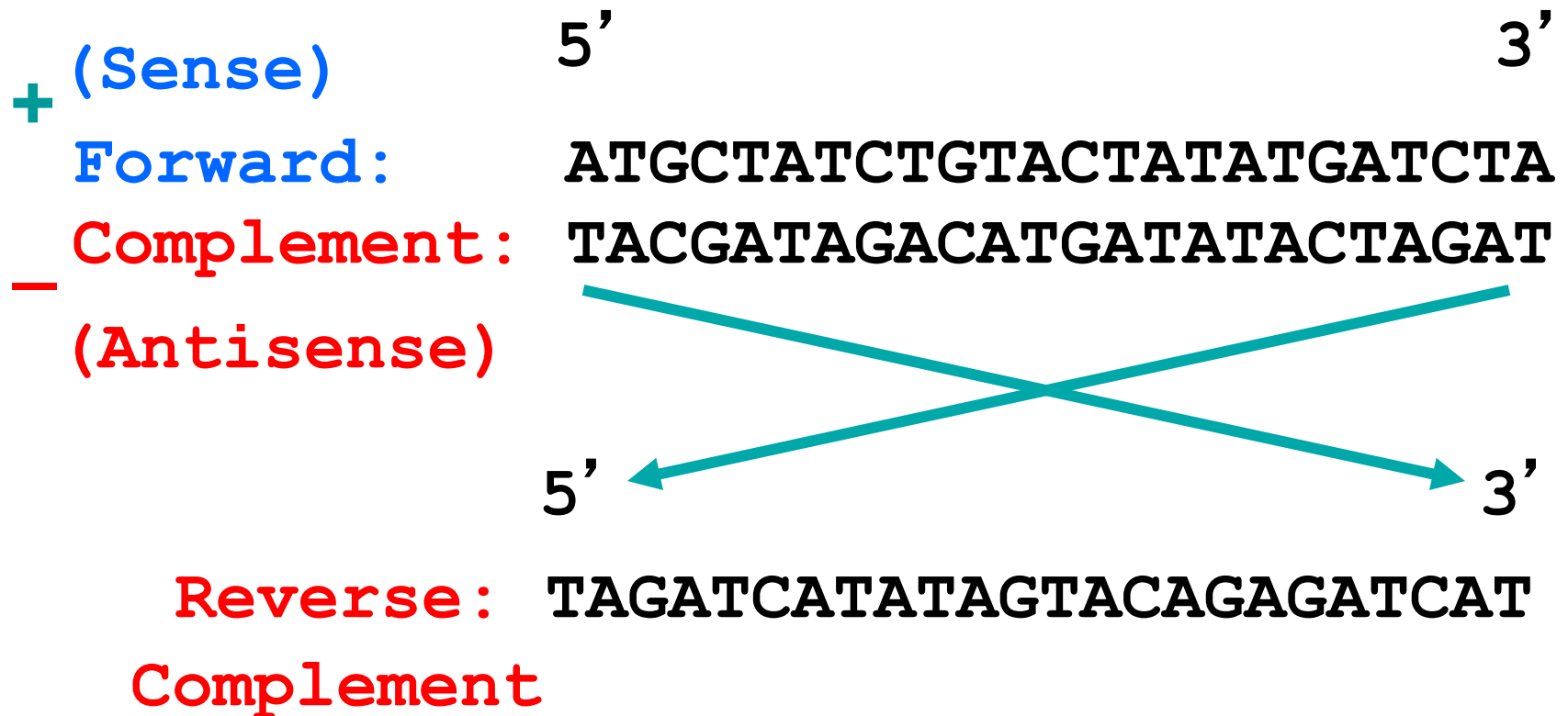
TACTAGCTATCTGACTAGCTAGCTAGCTAATCTAGG

3'

5'

<----Read this way

# DNA Sequence Nomenclature\*



# The Fundamental Paradigm

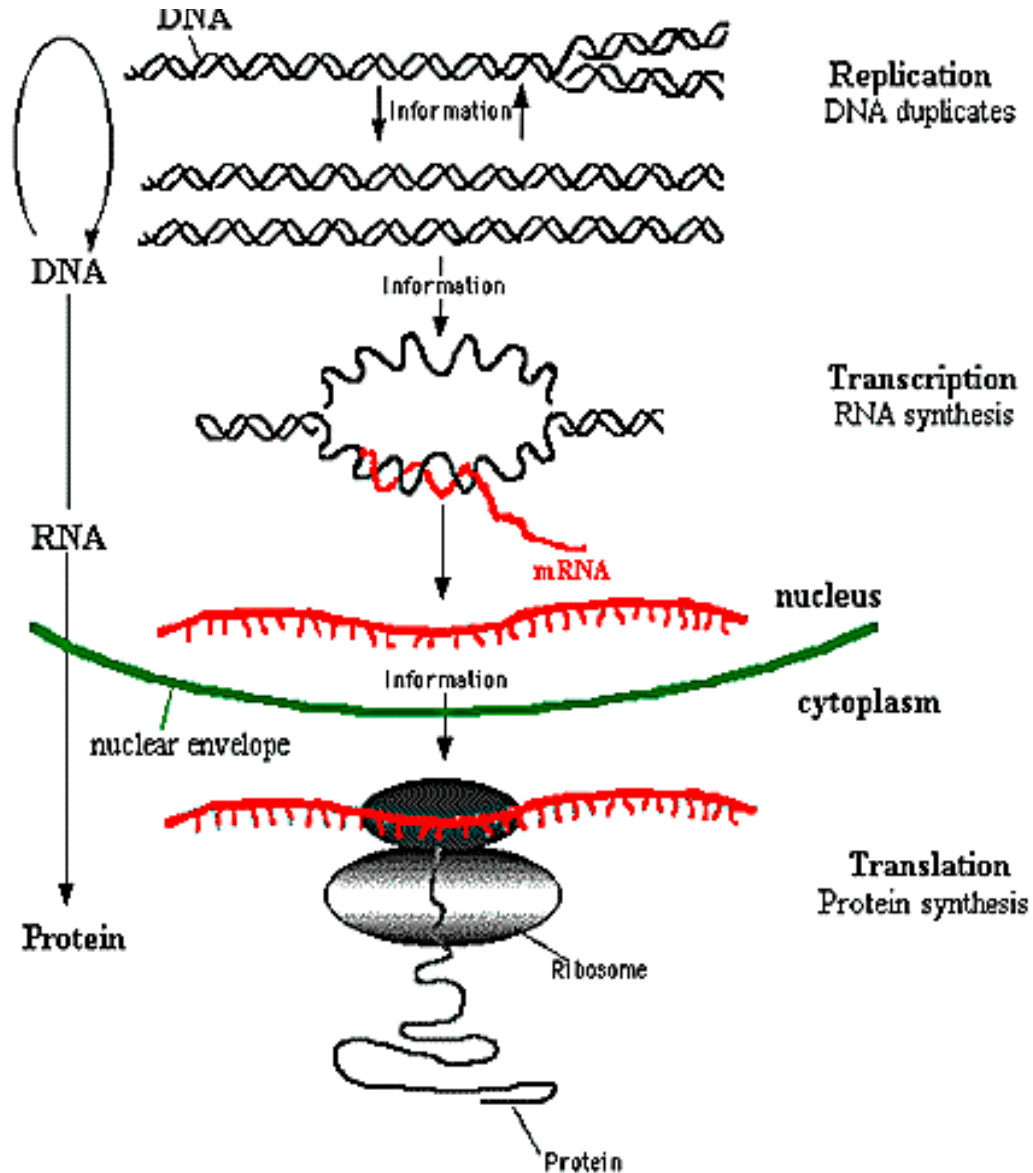
DNA



RNA



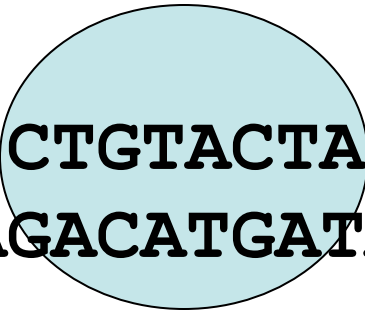
Protein



# RNA Polymerase

5' 3'

**Forward:** ATGCTATCTGTACTATATGATCTA  
**Complement:** TACGATAGACATGATATACTAGAT

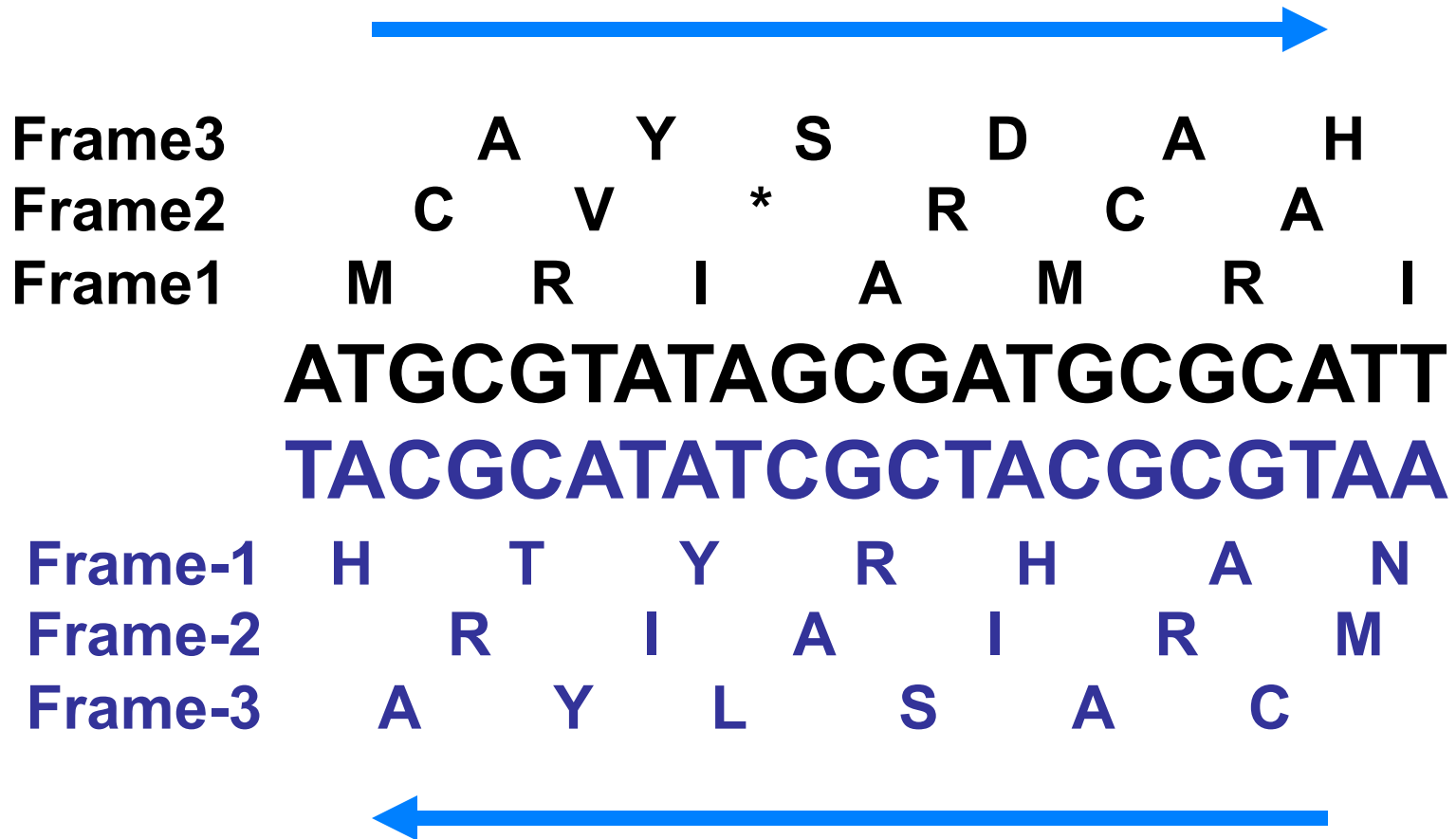
**Forward:** AUGCUAU  CTGTACTATATGATCTA  
**Complement:** TACGATAGACATGATATACTAGAT

# The Genetic Code\*

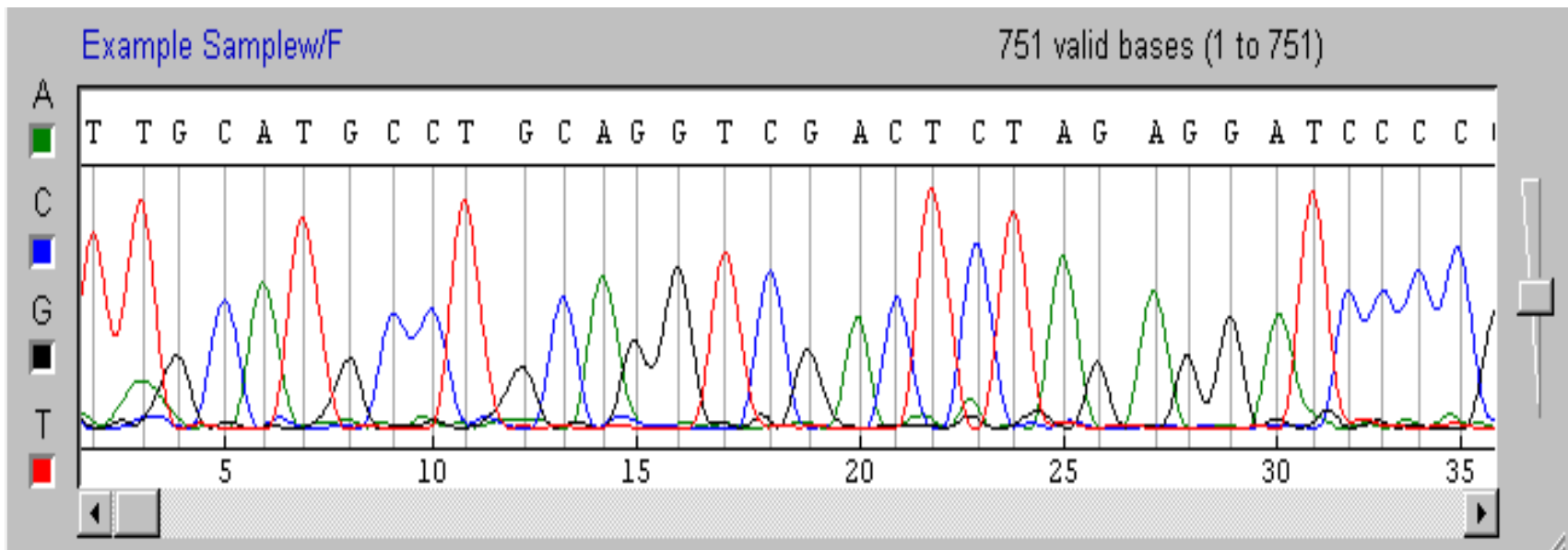
		SECOND BASE			
		U	C	A	G
FIRST BASE	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA } TERM UAG }	UGU } Cys UGC } UGA } TERM UGG } Trp
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }
	A	AUU } AUC }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



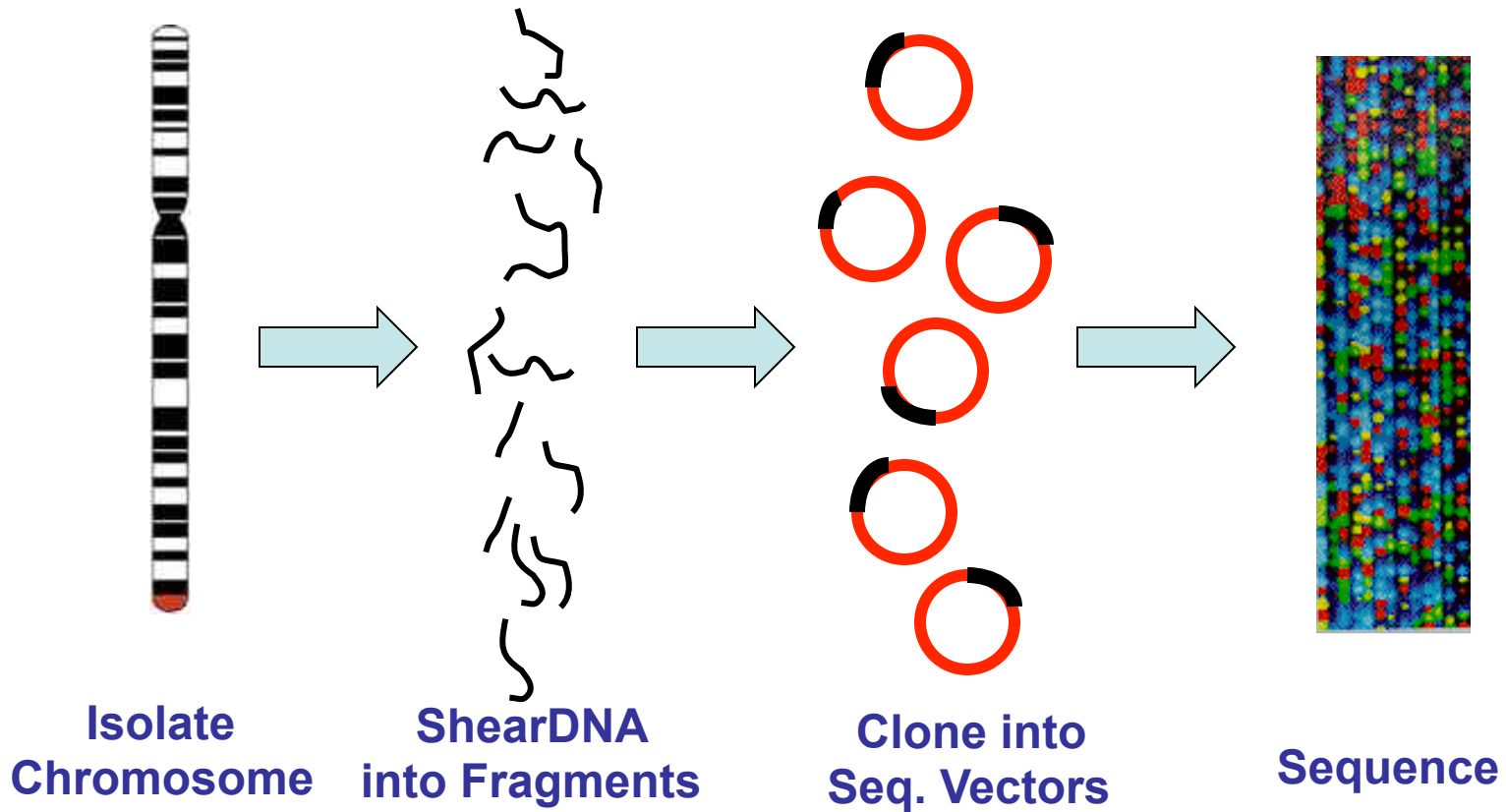
# Translating DNA/RNA\*



# DNA Sequencing



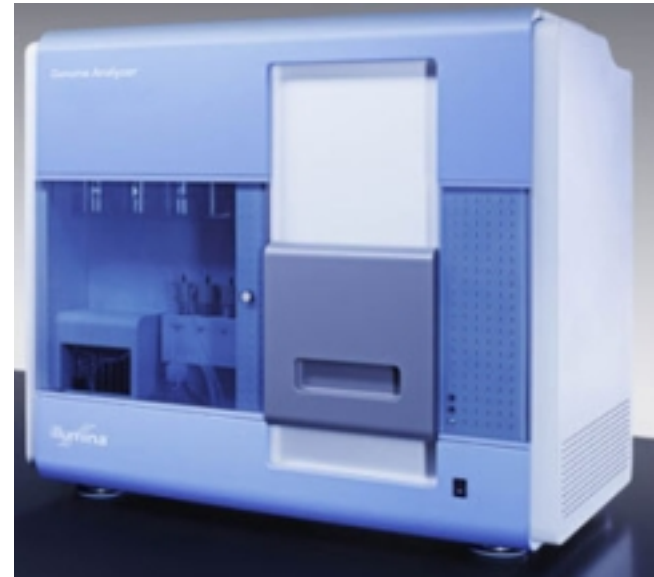
# Shotgun Sequencing\*



# Next Gen DNA Sequencing

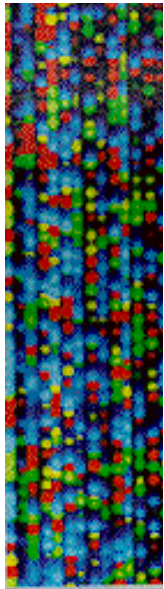


**ABI SOLiD - 20 billion bases/run  
Sequencing by ligation**



**Illumina/Solexa 15 billion bases/run  
Sequencing by dye termination**

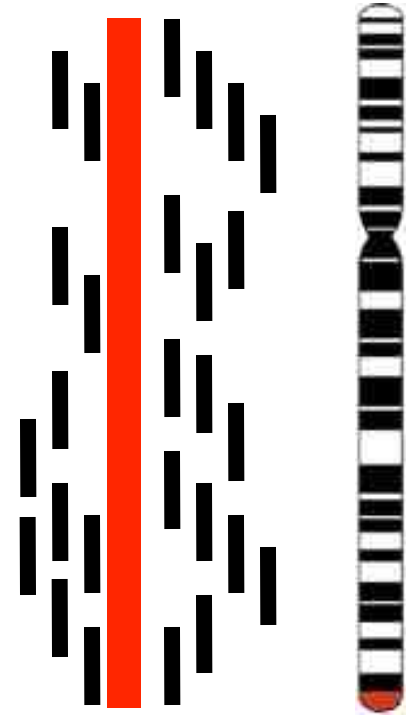
# Shotgun Sequencing



Sequence  
Chromatogram



Send to Computer



Assembled  
Sequence

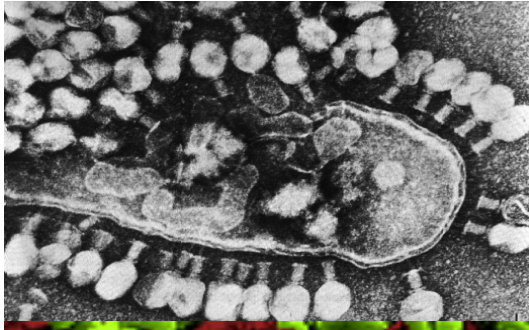
# Shotgun Sequencing

- Very efficient process for small-scale (~10 kb) sequencing (preferred method)
- First applied to whole genome sequencing in 1995 (*H. influenzae*)
- Now standard for all prokaryotic genome sequencing projects
- Successfully applied to *D. melanogaster*
- Moderately successful for *H. sapiens*

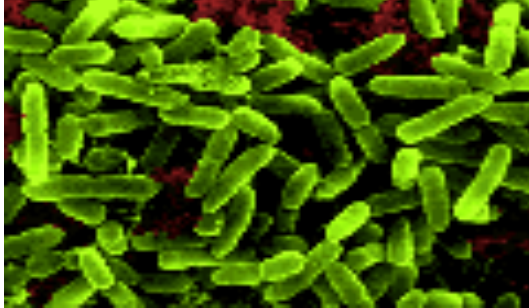
# The Finished Product

**GATTACAGATTACAGATTACAGATTACAGATTACAG  
ATTACAGATTACAGATTACAGATTACAGATTACAGA  
TTACAGATTACAGATTACAGATTACAGATTACAGAT  
TACAGATTAGAGATTACAGATTACAGATTACAGATT  
ACAGATTACAGATTACAGATTACAGATTACAGATTA  
CAGATTACAGATTACAGATTACAGATTACAGATTAC  
AGATTACAGATTACAGATTACAGATTACAGATTACA  
GATTACAGATTACAGATTACAGATTACAGATTACAG  
ATTACAGATTACAGATTACAGATTACAGATTACAGA  
TTACAGATTACAGATTACAGATTACAGATTACAGAT**

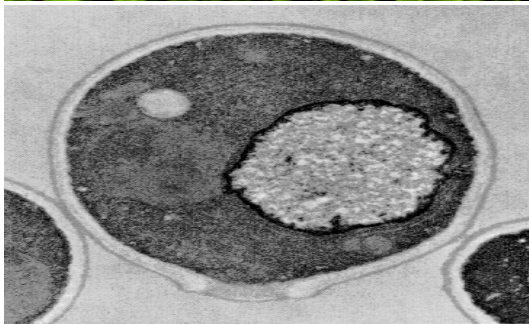
# Sequencing Successes\*



**T7 bacteriophage**  
completed in 1983  
39,937 bp, 59 coded proteins



**Escherichia coli**  
completed in 1998  
4,639,221 bp, 4293 ORFs



**Saccharomyces cerevisiae**  
completed in 1996  
12,069,252 bp, 5800 genes



# Sequencing Successes\*



**Caenorhabditis elegans**  
completed in 1998  
95,078,296 bp, 19,099 genes



**Drosophila melanogaster**  
completed in 2000  
116,117,226 bp, 13,601 genes



**Homo sapiens**  
completed in 2003  
3,201,762,515 bp, ~23,000 genes

# Genomes to Date

- **39 vertebrates** (human, mouse, rat, zebrafish, pufferfish, chicken, dog, chimp, cow, opossum)
- **35 plants** (arabidopsis, rice, poplar, corn, grape)
- **41 insects** (fruit fly, mosquito, honey bee, silkworm)
- **6 nematodes** (*C. elegans*, *C. briggsae*)
- **1 sea squirt**
- **32 parasites/protists** (plasmodium, guillardia)
- **54 fungi** (*S. cerevisiae*, *S. pombe*, *Aspergillus*)
- **3500+ bacteria and archeobacteria**
- **6000+ viruses**

<http://genomesonline.org/>

# Tracking Genomes

List of sequenced eukaryotic genomes - Wikipedia, the free encyclopedia

http://en.wikipedia.org/wiki/List\_of\_sequenced\_eukaryotic\_genomes

Department of Biology | Login - Department of Alberta | Audiobaba Music Search | Bioinformatics at the U of A | Coilgun Basics 2 | Pathguide: taxonomy resource list

[Cite this page](#)

## Chromista [edit]

The **Chromista** are a group of **protists** that contains the algal phyla **Heterokontophyta**, **Haptophyta** and **Cryptophyta**. Members of this group are mostly studied for evolutionary interest.

Organism <span>[x]</span>	Type <span>[x]</span>	Relevance <span>[x]</span>	Genome size <span>[x]</span>	Number of genes predicted <span>[x]</span>	Organization <span>[x]</span>	Year of completion <span>[x]</span>
<i>Guillardia theta</i>	Cryptomonad	Model organism	0.551 Mb (nucleomorph genome only)	464 <sup>[1]</sup>	Canadian Institute of Advanced Research, Philipps-University Marburg and the University of British Columbia	2001 <sup>[1]</sup>
<i>Thalassiosira pseudonana</i> Strain:CCMP 1335	Diatom		2.5 Mb	11,242 <sup>[2]</sup>	Joint Genome Institute and the University of Washington	2004 <sup>[2]</sup>
<i>Phaeodactylum tricomutum</i> Strain:CCAP1055/1	Diatom		27.4 Mb	10,402	Joint Genome Institute	2008 <sup>[3]</sup>

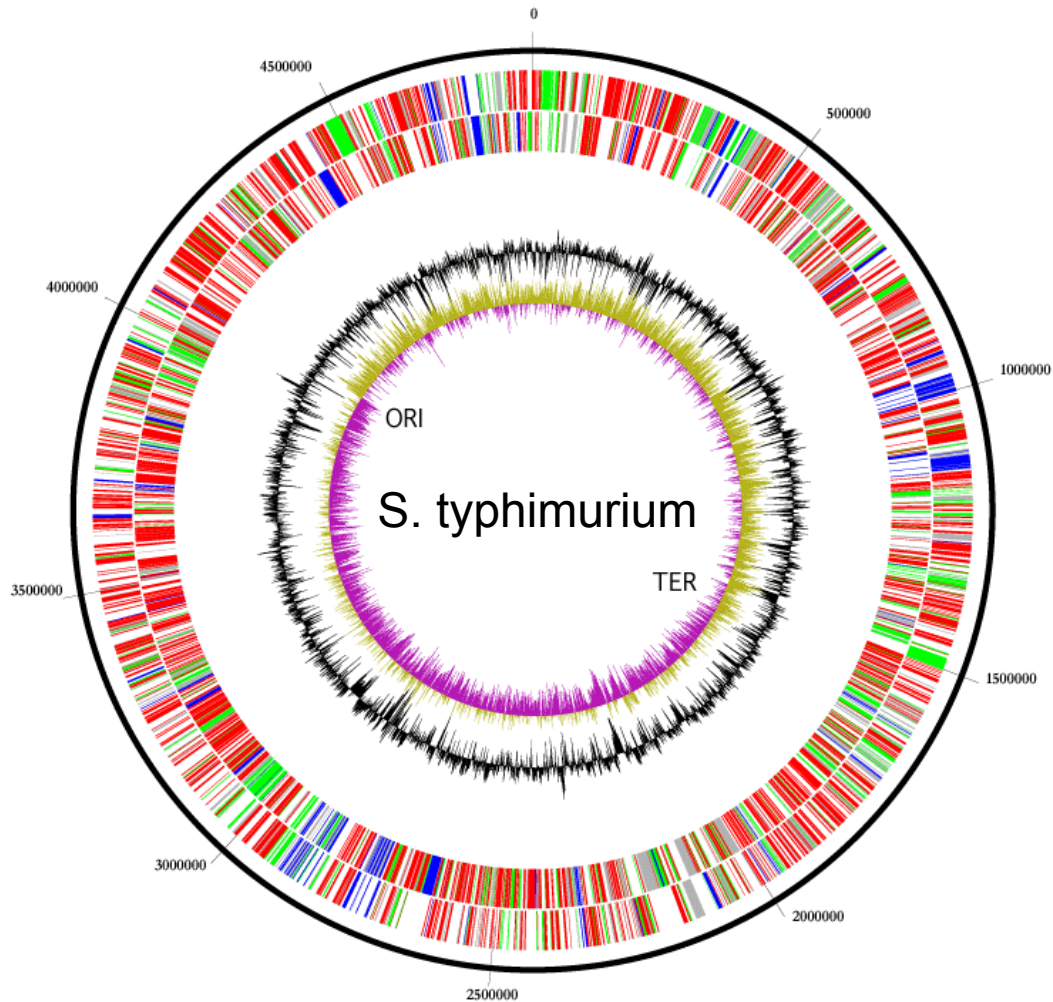
## Alveolata [edit]

**Alveolata** are a group of protists which includes the **Ciliophora**, **Apicomplexa** and **Dinoflagellata**. Members of this group are of particular interest to science as the cause of serious human and livestock diseases.

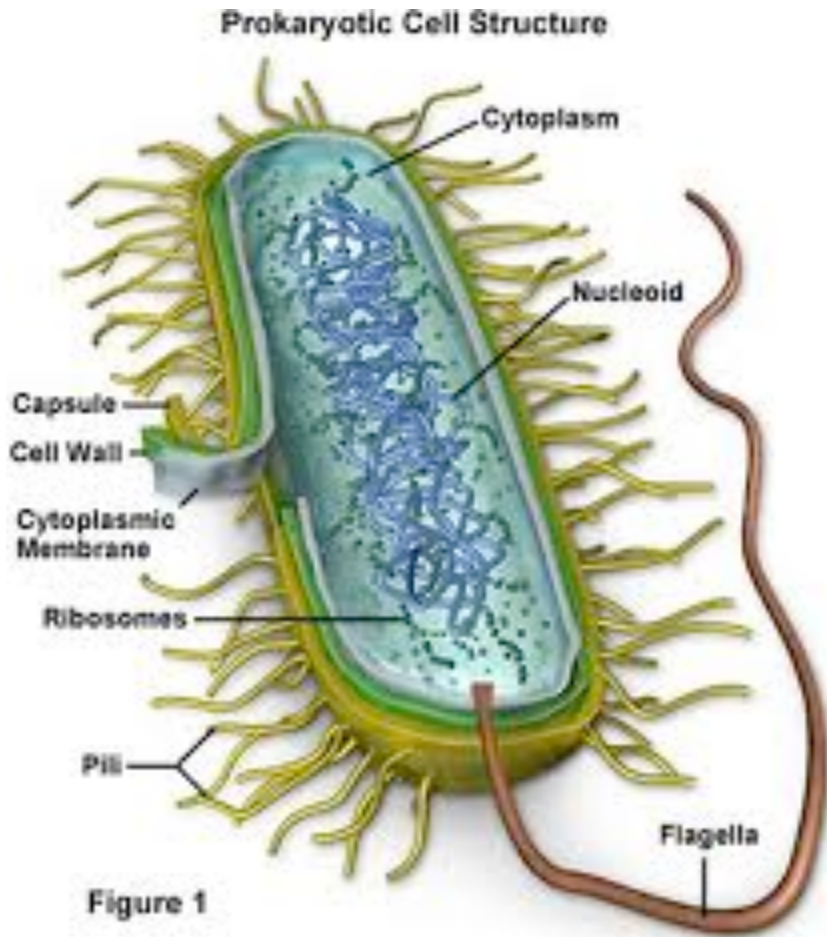
Organism <span>[x]</span>	Type <span>[x]</span>	Relevance <span>[x]</span>	Genome size <span>[x]</span>	Number of genes predicted <span>[x]</span>	Organization <span>[x]</span>	Year of completion <span>[x]</span>
<i>Babesia bovis</i>	Parasitic protozoan	Cattle pathogen	8.2 Mb	3,671		2007 <sup>[4]</sup>
<i>Cryptosporidium hominis</i> Strain:TU502	Parasitic protozoan	Human pathogen	10.4 Mb	3,994 <sup>[5]</sup>	Virginia Commonwealth University	2004 <sup>[5]</sup>
<i>Cryptosporidium parvum</i> C- or genotype 2 isolate	Parasitic protozoan	Human pathogen	16.5 Mb	3,807 <sup>[6]</sup>	UCSF and University of Minnesota	2004 <sup>[6]</sup>
<i>Paramecium tetraurelia</i>	Ciliate	Model organism	72 Mb	39,642 <sup>[7]</sup>	Genoscope	2008 <sup>[7]</sup>

[http://en.wikipedia.org/wiki/List\\_of\\_sequenced\\_eukaryotic\\_genomes](http://en.wikipedia.org/wiki/List_of_sequenced_eukaryotic_genomes)

# Gene Finding in Prokaryotes



# Prokaryotes

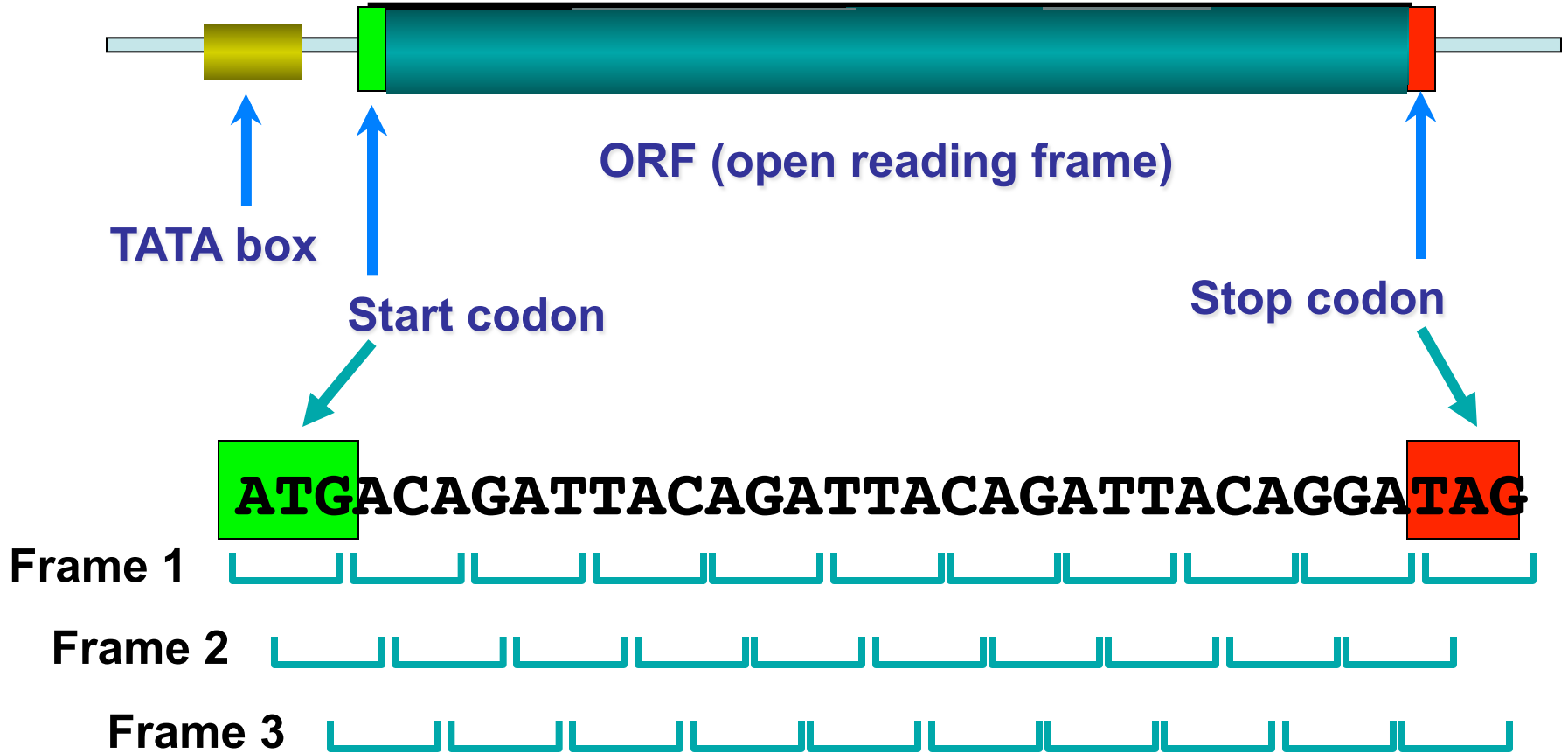


- Are a group of unicellular organisms whose cells lack a cell nucleus (karyon), or any other membrane-bound organelles
- Divided into bacteria and archaea

# Prokaryotes\*

- **Simple gene structure**
- **Small genomes (0.5 to 10 million bp)**
- **No introns (uninterrupted)**
- **Genes are called Open Reading Frames of “ORFs” (include start & stop codon)**
- **High coding density (>90%)**
- **Some genes overlap (nested)**
- **Some genes are quite short (<60 bp)**

# Prokaryotic Gene Structure\*



# Gene Finding In Prokaryotes\*

- **Scan forward strand until a start codon is found**
- **Staying in same frame scan in groups of three until a stop codon is found**
- **If # of codons between start and end is greater than 50, identify as gene and go to last start codon and proceed with step 1**
- **If # codons between start and end is less than 50, go back to last start codon and go to step 1**
- **At end of chromosome, repeat process for reverse complement**



# ORF Finding Tools

- <http://www.ncbi.nlm.nih.gov/gorf/gorf.html>
- [http://www.bioinformatics.org/sms2/orf\\_find.html](http://www.bioinformatics.org/sms2/orf_find.html)
- <https://www.dna20.com/toolbox/ORFFinder.html>
- <http://www0.nih.go.jp/~jun/cgi-bin/frameplot.pl>

# NCBI ORF Finder

**ORF Finder (Open Reading Frame Finder)**

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI  
for data mining

GenBank  
sequence submission support and software

FTP site  
download data and software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence in a sequence already in the database.

This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION   
or sequence in FASTA format

OrfFind Clear

Document: Done

<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>

# Type in or Paste DNA Sequence

ORF Finder

www.ncbi.nlm.nih.gov/gorf/gorf.html

NCBI

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI  
Tools for data mining  
GenBank  
sequence submission support and software  
FTP site  
download data and software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION  Orffind Clear

or sequence in FASTA format

```
>sequence
ATGGCGTAGCGTGATCGATGCTAGTTTAGCCGAGCTACGACTATTCTATACGGAC
TAGCGATCGACTAGCATCGACACTACTACTAGATGATAGTATCTACTCGACTCA
TCTCACTGAAGTATTAGTAATTAATGGCGTAGCGTGATCGATGCTAGTTAGCCG
AGCTACGACTATTCTATACGGACTAGCGATCGACTAGCATCGACACTACTATCTA
GATGATAGTATCTAGTCGACTCATCTCACTGAAGTATTAGTAATTAATGGCGTAG
CGTGATCGATGCTAGTTTAGCCGAGCTACGACTATTCTATACGGACTAGCGATCG
ACTAGCATCGACACTACTATCTAGATGATAGTATCTAGTCGACTCATCTCACTGA
AGTATTAGTAATTAATGGCGTAGCGTGATCGATGCTAGTTTAGCCGAGCTACGAC
```

FROM:  TO:

Genetic codes 1 Standard

Comments and suggestions to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov)  
Credits to: [Tatiana Tatusov](#) and [Roman Tatusov](#)

Press "Orffind"

# NCBI ORF Finder

ORF Finder

www.ncbi.nlm.nih.gov/gorf/orfig.cgi

ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxonomy Structure

sequence

View 1 GenBank Redraw 100 SixFrames FramefromtoLength

Click Six frames button

# NCBI ORF Finder

ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxonomy Structure

sequence

View 3 Fasta protein ViewAll Redraw OrfFind

Press GenBank button to toggle to Fasta protein format

Click on any of the 6 marked “bars” to view any of the 6 reading frames

# NCBI ORF Finder

The screenshot shows the NCBI ORF Finder web interface. At the top, the title "ORF Finder (Open Reading Frame Finder)" is displayed. Below the title is a navigation bar with links to PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The "sequence" section contains a form with "Program" set to "blastp", "Database" set to "nr", and a "BLAST" button. A "Cognitor" button is also present. Below the form, there are buttons for "View", "2 Fasta nucleotide", "ViewAll", "Redraw", and "OrfFind". The main content area displays a sequence analysis result with a protein length of 176 aa. The sequence is shown in a multi-line format with amino acid abbreviations and their corresponding codons. The sequence is: 2 tggcgtagcgtgacgatcgtagtttagccgagctacgactattctatacggactagcga  
W R S V I D A S L A E L R L F Y T D \* R  
62 tcgactagcgcacactactatcagatgatagatctagtgactcctcactgaag  
S T S I D T T I \* M I V S S R L I S L K  
122 tattagtaataatggcglagcgtgacgatcgttagttagccgagctacgactattcta  
Y \* \* L M A \* R D R C \* F S R A T T I L  
182 tacggactagcgcacactactatcagatgatagatcagtcgact  
Y G L A I D \* H R H Y Y L D D S I \* S T  
242 catctcactgaagtattagtaataatggcgtagcglagcgcgtagtttagccgagc  
H L T E V L V I N G V A \* S M L V \* P S  
302 tacgactattctatacggactagcgcacactactatcagatgatagatcagatg  
Y D Y S I R T S D R L A S T L L S R \* \*  
362 tatctagtcgactcctcactcagtagtattagtaataatggcgtagcgtgacgatcgt  
Y L V D S S H \* S I S N \* W R S V I D A

# Using Other ORF Finders

- **Go to the website**
- **Paste in some random DNA sequence or use the example sequence provided on the website**
- **Press the submit button**
- **Output will typically be displayed in a pop-up window showing the translation of the protein(s)**

# But...

- **Prokaryotic genes are not always so simple to find**
- **When applied to whole genomes, simple ORF finding programs tend to overlook small genes and tend to overpredict the number of long genes**
- **Can we include other genome signals?**
- **Can we account for alternative start and stop signals?**



# Key Prokaryotic Gene Signals\*

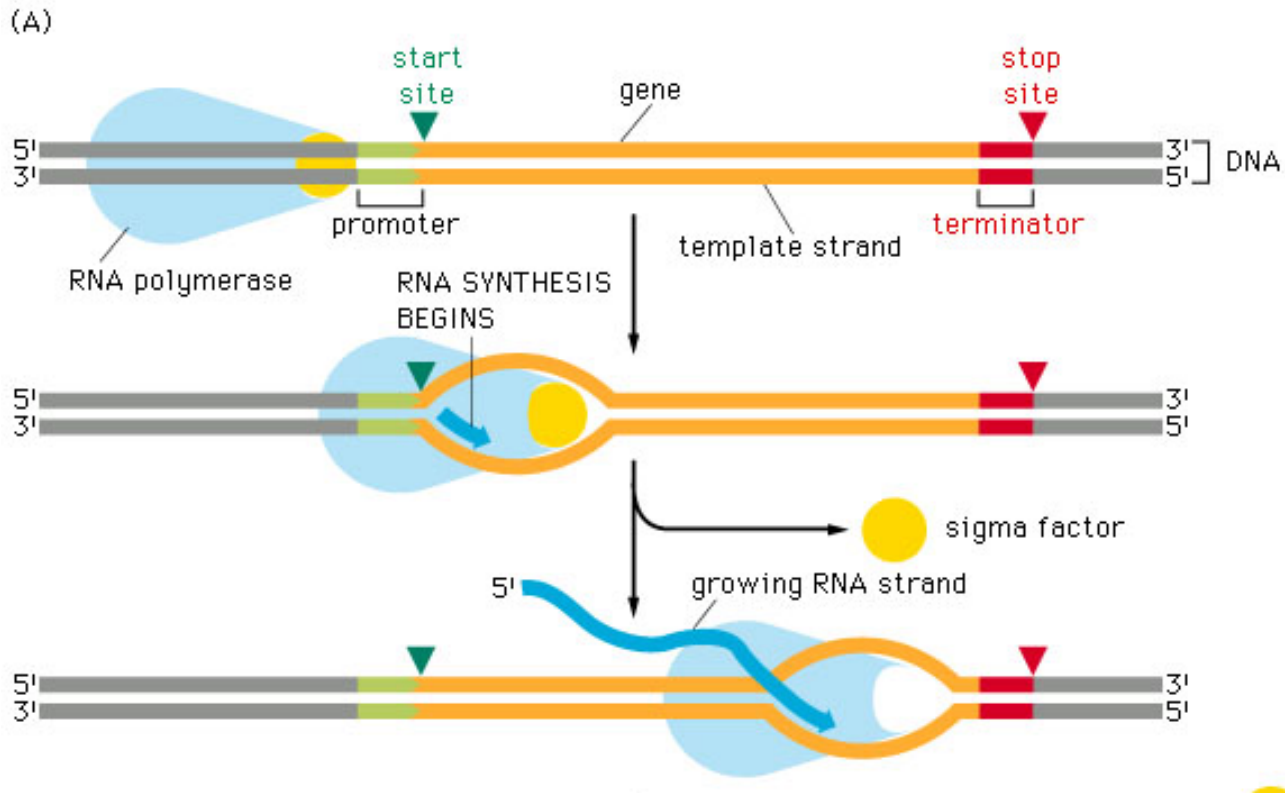
- **Alternate start codons**
- **RNA polymerase promoter site (-10, -35 site or Pribnow box)**
- **Shine-Dalgarno sequence (Ribosome binding site-RBS)**
- **Stem-loop (rho-independent) terminators**
- **High GC content (CpG islands)**

# Alternate Start Codons (E. coli)

<b>Class I</b>	<b>ATG</b>	<b>Met</b>
	<b>GTG</b>	<b>Val</b>
	<b>TTG</b>	<b>Leu</b>
<b>Class IIa</b>	<b>CTG</b>	<b>Met</b>
	<b>ATT</b>	<b>Val</b>
	<b>ATA</b>	<b>Leu</b>
	<b>ACG</b>	<b>Thr</b>

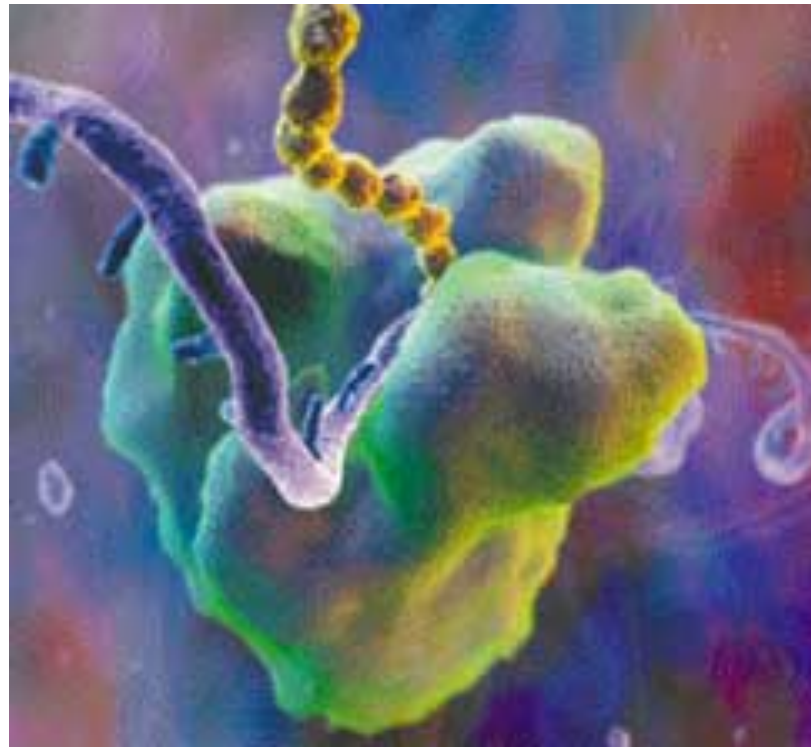
# -10, -35 Site (RNA pol Promoter)

-36 -35 -34 -33 -32 ... -12 -11 -10 -9 -8 -7  
T T G A C ... T A t A A T



# RBS (Shine Dalgarno Seq)

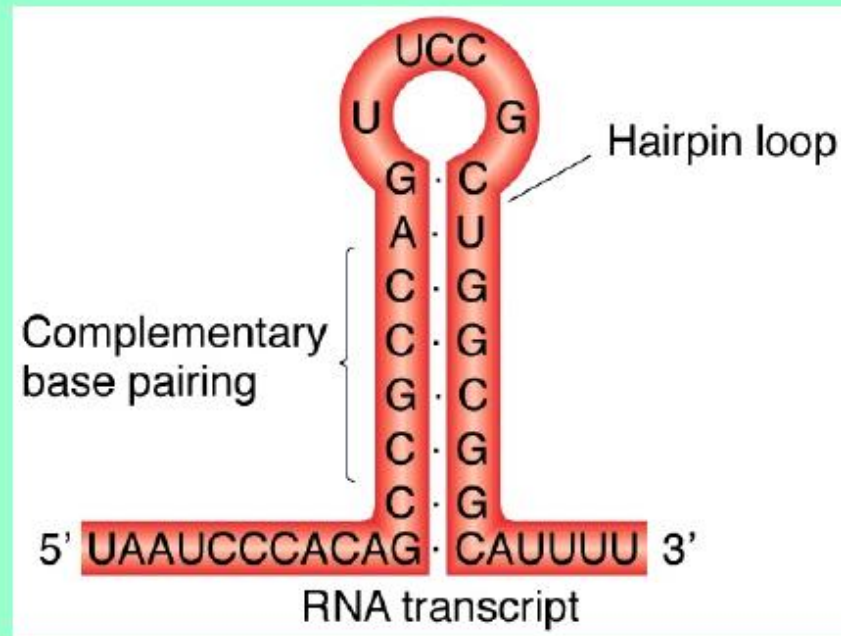
-17 -16 -15 -14 -13 -12 .. -1 0 1 2 3 4  
A G G A G G n **A** **T** **G** n C



Recruits bacterial ribosome to bind the mRNA strand

# Terminator Stem-loops

rho-independent terminator



# A Better Gene Finder...

- **Scan for ORFs using regular and alternate codons**
- **Among the ORFs found, check for RNA Pol promoter sites and RBS binding sites on 5' end – if found, keep the ORF**
- **Among the ORFs found look for stem-loop features – if found, keep the ORF**
- **How best to find these extra signals or signal sites?**

# Simple Methods to Gene Site Identification\*



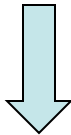
**A PSSM**

- Use a consensus sequence (CNNTGA)
- Use a regular expression (C[**TG**]A\*)
- Use a custom scoring matrix called a position specific scoring matrix (PSSM) built from multiple sequence alignments

# Building a PSSM - Step 1\*

A T T T A G T A T C  
G T T C T G T A A C  
A T T T T G T A G C  
A A G C T G T A A C  
C A T T T G T A C A

*Multiple  
Alignment*



<b>A</b>	3	2	0	0	1	0	0	5	2	1
<b>C</b>	1	0	0	2	0	0	0	0	1	4
<b>G</b>	1	0	1	0	0	5	0	0	1	0
<b>T</b>	0	3	4	3	4	0	5	0	1	0

*Table of  
Occurrences*



# Building a PSSM - Step 2\*

<b>A</b>	3	2	0	0	1	0	0	5	2	1
<b>C</b>	1	0	0	2	0	0	0	0	1	4
<b>G</b>	1	0	1	0	0	5	0	0	1	0
<b>T</b>	0	3	4	3	4	0	5	0	1	0

*Table of Occurrences*



<b>A</b>	.6	.4	0	0	.2	0	0	1	.4	.2
<b>C</b>	.2	0	0	.4	0	0	0	0	.2	.8
<b>G</b>	.2	0	.2	0	0	1	0	0	.2	0
<b>T</b>	0	.6	.8	.6	.8	0	1	0	.2	0

*PSSM with no pseudocounts*

# Pseudocounts\*

- **Method to account for small sample size of multi-sequence alignment**
- **Gets around problem of having “0” score in PSSM or profile**
- **Defined by a correction factor “B” which reflects overall composition of sequences under consideration**
- **$B = \sqrt{N}$  or  $B = 0.1$  which falls off with N where  $N = \#$  sequences**

# Pseudocounts\*

- **Score( $X_i$ ) =  $(q_x + p_x)/(N + B)$**
- **q = observed counts of residue X at pos. i**
- **p = pseudocounts of X =  $B \cdot \text{frequency}(X)$**
- **N = total number of sequences in MSA**
- **B = number of pseudocounts (assume  $\sqrt{N}$ )**

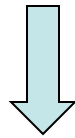
$$\text{Score}(A_1) = (3 + \sqrt{5}(0.32)) / (5 + \sqrt{5}) = 0.51$$

0.32 is the frequency of A's over the entire genome sequence

# Including Pseudocounts - Step 2\*

<b>A</b>	3	2	0	0	1	0	0	5	2	1
<b>C</b>	1	0	0	2	0	0	0	0	1	4
<b>G</b>	1	0	1	0	0	5	0	0	1	0
<b>T</b>	0	3	4	3	4	0	5	0	1	0

*Table of  
Occurrences*



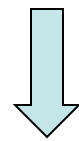
<b>A</b>	.51	.38	.09	.09	.24	.09	.09	.79	.38	.24
<b>C</b>	.19	.06	.06	.33	.06	.06	.06	.06	.19	.61
<b>G</b>	.19	.06	.19	.06	.06	.75	.06	.06	.19	.06
<b>T</b>	.09	.51	.65	.51	.65	.09	.79	.09	.24	.09

*PSSM with  
pseudocounts*

# Calculating Log-odds - Step 3\*

<b>A</b>	.51	.38	.09	.09	.24	.09	.09	.79	.38	.24
<b>C</b>	.19	.06	.06	.33	.06	.06	.06	.06	.19	.61
<b>G</b>	.19	.06	.19	.06	.06	.75	.06	.06	.19	.06
<b>T</b>	.09	.51	.65	.51	.65	.09	.79	.09	.24	.09

*PSSM with  
pseudocounts*



$-\text{Log}_{10}$

<b>A</b>	0.2	0.4	1.1	1.1	0.7	1.1	1.1	0.1	0.4	0.7
<b>C</b>	0.7	1.2	1.2	0.4	1.2	1.2	1.2	1.2	0.7	0.1
<b>G</b>	0.7	1.2	0.7	1.2	1.2	0.1	1.2	1.2	0.7	1.2
<b>T</b>	1.1	0.2	0.1	0.2	0.1	1.1	0.1	1.1	0.7	1.1

*Log-odds  
PSSM*

# Scoring a Sequence - Step 4\*

<b>A</b>	0.2	0.4	1.1	1.1	0.7	1.1	1.1	0.1	0.4	0.7
<b>C</b>	0.7	1.2	1.2	0.4	1.2	1.2	1.2	1.2	0.7	0.1
<b>G</b>	0.7	1.2	0.7	1.2	1.2	0.1	1.2	1.2	0.7	1.2
<b>T</b>	1.1	0.2	0.1	0.2	0.1	1.1	0.1	1.1	0.7	1.1

**Log-odds  
PSSM**

**A T T T A G T A T C**

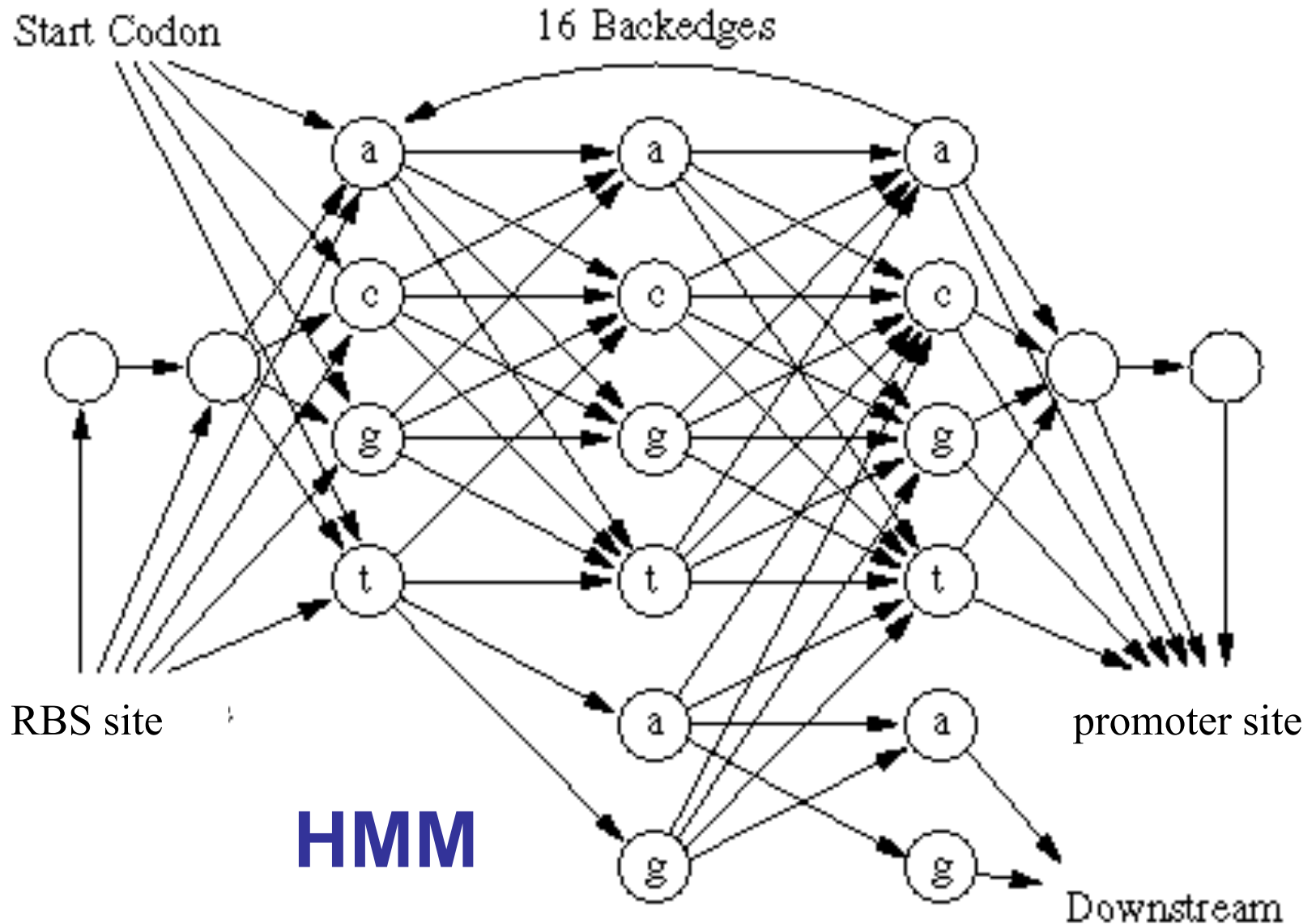
**Score = 2.5**  
*(Lowest score wins)*

<b>A</b>	0.2	0.4	1.1	1.1	0.7	1.1	1.1	0.1	0.4	0.7
<b>C</b>	0.7	1.2	1.2	0.4	1.2	1.2	1.2	1.2	0.7	0.1
<b>G</b>	0.7	1.2	0.7	1.2	1.2	0.1	1.2	1.2	0.7	1.2
<b>T</b>	1.1	0.2	0.1	0.2	0.1	1.1	0.1	1.1	0.7	1.1

# How to Use a PSSM

- **Specific PSSMs can be made for finding RNA Pol promoter sites and RBS binding sites as well as many eukaryotic signal sites**
- **PSSMs can also be made for finding stem loop structures and other genetic features**
- **Sort of “custom” BLOSUM scoring matrices like those used in BLAST**
- **Very popular in the 1980s-1990s**

# More Sophisticated Methods





# Hidden Markov Models

- **Special kind of machine learning (artificial intelligence) method that is often used in pattern recognition problems such as speech recognition (Siri, Dragon NaturallySpeaking), handwriting recognition, gesture recognition, part-of-speech tagging, musical score following and bioinformatics**

# More Sophisticated Prokaryotic Gene Finding Methods

- **GLIMMER 3.0**
  - <http://cbcb.umd.edu/software/glimmer/>
  - Uses interpolated markov models (IMM)
  - Requires training of sample genes
  - Takes about 1 minute/genome
- **GeneMark.hmm**
  - [http://opal.biology.gatech.edu/GeneMark/gmhmm2\\_prok.cgi](http://opal.biology.gatech.edu/GeneMark/gmhmm2_prok.cgi)
  - Available as a web server
  - Uses hidden markov models (HMM)

# Glimmer 3.02 Website

The screenshot shows a Mozilla Firefox browser window displaying the Glimmer 3.02 website. The address bar shows the URL [http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer\\_3.cgi](http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi). The page features the NCBI logo and the title "Microbial Genomes" with a colorful hexagonal pattern background. A navigation menu includes links for HOME, SEARCH, SITE MAP, Genome Project, Genome, Prokaryotic Projects, Collaborators, gMap, ProtMap, TaxPlot, BLAST, FTP, and Contact us. The main content area is titled "Microbial Genome Annotation Tools" and contains a description of GLIMMER: "GLIMMER is a system for finding genes in microbial DNA, especially the genomes of bacteria and archaea. GLIMMER (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models to identify coding regions." Below this description are two bullet points listing publications by Delcher et al. (1999) and Salzberg et al. (1998). A "Download GLIMMER" link is provided, pointing to the Center for Bioinformatics and Computational Biology. On the right side, there is a vertical menu with sections: Genomes (containing links for Genome Projects, Prokaryotic Projects, Microbial Genomes, Home, Complete Genomes, Draft Assemblies, Registered, Plasmids, Entrez Genome), Submit a Genome, Sequin, Submission Guide, Register a Project, Submit a Genome, Submit Traces, Tools, Resources, Sequencing Centers, Collaborators, and Statistics. At the bottom, there is a form for uploading a sequence from a file (with a "Browse..." button) and a text area for pasting a FASTA sequence.

[http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer\\_3.cgi](http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi)

# Glimmer Performance

## *Glimmer 2.0's Accuracy*

Organism	Genes annotated	Annotated genes found	% found
H. influenzae	1738	1720	99.0
M. genitalium	483	480	99.4
M. jannaschii	1727	1721	99.7
H. pylori	1590	1550	97.5
E. coli	4269	4158	97.4
B. subtilis	4100	4030	98.3
A. fulgidis	2437	2404	98.6
B. burgdorferi	853	843	99.3
T. pallidum	1039	1014	97.6
T. maritima	1877	1854	98.8

# Genemark.hmm

GeneMark.hmm for Prokaryotes

opal.biology.gatech.edu/Genemark/gmhmm2\_prok.cgi

Google

## GeneMark.hmm for Prokaryotes (Version 2.8) [\(Reload this page\)](#)

**Reference:** Lukashin A. and Borodovsky M., [GeneMark.hmm: new solutions for gene finding](#), **NAR**, 1998, Vol. 26, No. 4, pp. 1107-1115.  
[\[ Download PDF \]](#)

Prediction models have been pre-computed for a [265](#) completely sequenced prokaryotic genomes from the NCBI RefSeq database. Gene predictions made for these genomes are available in the [GeneMark prokaryotic database](#).

### Input Sequence

Title (optional):

Sequence Text:

Sequence File upload:

Species:

Use RBS model, if available

### Output Options

E-Mail Address (required for graphical output or sequences longer than 5000000 bp)

# EasyGene (A Late Entry)

EasyGene 1.2 Server

www.cbs.dtu.dk/services/EasyGene/

Shine Dalgarno PSSM

CBS >> CBS Prediction Servers >> EasyGene

## EasyGene 1.2b Server

### Gene finding in prokaryotes

The EasyGene 1.2 server produces a list of predicted genes given a sequence of prokaryotic DNA. The current version contains models for [138 different organisms](#). Each prediction is attributed with a significance score (R-value) indicating how likely it is to be just a non-coding open reading frame rather than a real gene. All that is required of you as a user is to supply the query sequence(s) and to select the organism model to use (see [instructions](#)).

The pre-calculated EasyGene 1.2 predictions for the complete genomes of the 138 organisms can be downloaded from the [EasyGene site](#) at [BINP](#) at the University of Copenhagen.

This version replaces EasyGene 1.0. View the [version history](#) of this server.

[Instructions](#)      [Output format](#)      [Article abstracts](#)

### SUBMISSION

Paste a single sequence or several sequences in [FASTA](#) format into the field below:

Submit a file in [FASTA](#) format directly from your local disk:

Organism:       View the [organism list](#).

R-value cutoff:        Predict suboptimal gene starts

**Restrictions:**  
At most 10,000,000 nucleotides per submission in at most 50 sequences.

**Confidentiality:**  
The sequences are kept confidential and will be deleted after processing.

CITATIONS

TECHNICAL UNIVERSITY OF DENMARK DTU

<http://www.cbs.dtu.dk/services/EasyGene/>

# EasyGene Output

**DESCRIPTION**

The output conforms to the [GFF](#) format. For each input sequence the server prints a list of predicted genes, one per line. The columns are:

- **seqname**: input sequence name;
- **model**: organism model code (also in plain text in the table head);
- **feature**: predicted feature, 'CDS' or 'CDSsub' (alternative translation start);
- **start** and **end**: positions in the sequence;
- **score**: R-value, indicating how likely the fragment is to be just a non-coding open reading frame rather than a real gene;
- **strand**: '+' or '-';
- **startc**: predicted start codon;
- **odds**: log odds score.

Only the predictions with R-values lower than the selected R-value cutoff (the default is 2) are reported.

The example below shows the EasyGene 1.2 output for the sequence taken from the GenBank entry [AB010576](#), containing *Bacillus subtilis* **ComX**, **ComQ** and **DegQ** genes. All the three genes are predicted as annotated in the database (shown in green), with high confidence, although an alternative translation start is preferred for **comQ** (shown in orange). Two additional genes not annotated in the GenBank entry are also predicted.

**EXAMPLE OUTPUT**

```
##gff-version 2
##source-version easygene-1.2b
##date 2007-08-15
##Type DNA
# model: BS03 Bacillus subtilis
# seqname      model  feature start  end    score      +/-    ?    startc  odds
# -----
AB010576      BS03   CDS      67    324    0.0271875  +     0    #ATG    20.1861
AB010576      BS03   CDSsub   55    324    0.031955   +     0    #ATG    20.1731
AB010576      BS03   CDS      1129  1269  0.0190622  +     0    #ATG    15.7102
AB010576      BS03   CDS      1370  2314  2.13273e-12 +     0    #ATG    74.7815
AB010576      BS03   CDSsub   1454  2314  1.92405e-12 +     0    #ATG    74.6356
AB010576      BS03   CDS      2327  2491  0.0167943  +     0    #ATG    17.2951
AB010576      BS03   CDS      300   668   1.43511    -     0    #ATG    10.6215
# -----
```

# Gene Finding with GLIMMER & Company

- **Go to your preferred website**
- **Paste in the DNA sequence of your favorite PROKARYOTIC genome (this won't work for eukaryotic genomes and it won't necessarily work for viral genomes, it may work for phage genomes)**
- **Press the submit button**
- **Output will typically be presented in a new screen or emailed to you**



# Bottom Line...\*

- **Gene finding in prokaryotes is now a “solved” problem**
- **Accuracy of the best methods approaches 99%**
- **Gene predictions should always be compared against a BLAST search to ensure accuracy and to catch possible sequencing errors**
- ***Homework: Try testing some of the web servers I have mentioned today***