Bioinformatics-II BIF-210
3(3+0)

Objective

To develop understanding of gene and protein at structural level

Dr. Rana Rehan Khalid
Assistant Professor
NCB, Qau, Islamabad

WHAT IS BIOINFORMATICS?

Luscombe et al. in defining bioinformatics as a union of biology and informatics:

- Bioinformatics involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins.

- The emphasis here is on the use of computers because most of the tasks in genomic data analysis are highly repetitive or mathematically complex. The use of computers is absolutely indispensable in mining genomes for information gathering and knowledge building.

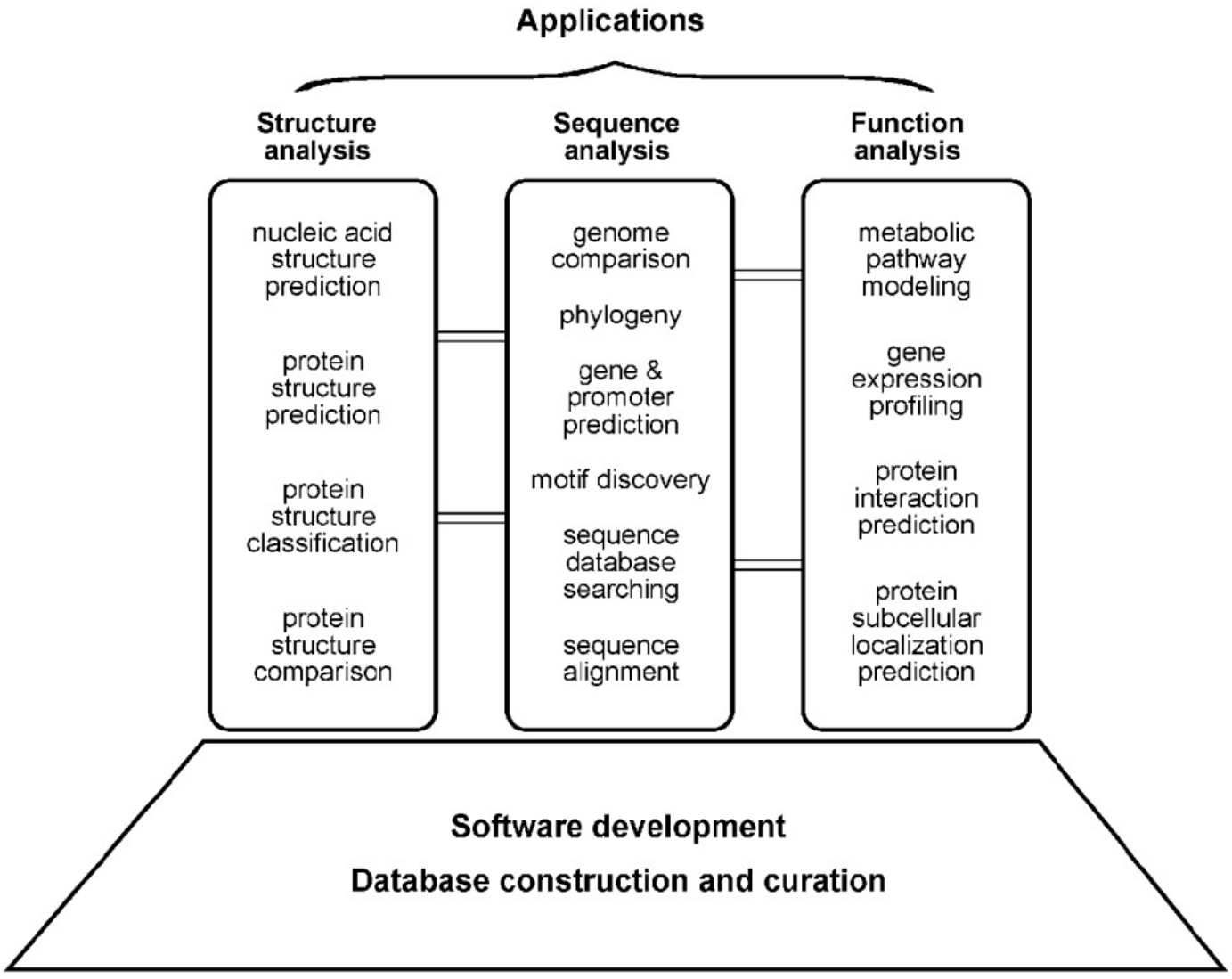
Bioinformatics differs from a related field known as computational biology

- Bioinformatics is limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products and is often considered computational molecular biology.

- Computational biology encompasses all biological areas that involve computation. For example, mathematical modeling of ecosystems, population dynamics, application of the game theory in behavioral studies, and phylogenetic construction using fossil records all employ computational tools, but do not necessarily involve biological macromolecules.

SCOPE

It is worth noting that there are other views of how the two terms relate. For example, one version defines bioinformatics as the development and application of computational tools in managing all kinds of biological data, whereas computational biology is more confined to the theoretical development of algorithms used for bioinformatics. The confusion at present over definition may partly reflect the nature of this vibrant and quickly evolving new field.


Applications
Structure analysis
Sequence analysis
Function analysis
nucleic acid structure prediction
protein structure prediction
protein structure classification
protein structure comparison
genome comparison
phylogeny
gene & promoter prediction
motif discovery
sequence database searching
sequence alignment
metabolic pathway modeling
gene expression profiling
protein interaction prediction
protein subcellular localization prediction
Software development
Database construction and curation

GOAL

The ultimate goal of bioinformatics is to better understand a living cell and how it functions at the molecular level. By analyzing raw molecular sequence and structural data, bioinformatics research can generate new insights and provide a “global” perspective of the cell.

WHY

The reason that the functions of a cell can be better understood by analyzing sequence data is ultimately because the flow of genetic information is dictated by the “central dogma” of biology in which DNA is transcribed to RNA, which is translated to proteins.

Cellular functions are mainly performed by proteins whose capabilities are ultimately determined by their sequences. Therefore, solving functional problems using sequence and sometimes structural approaches has proved to be a fruitful endeavor.

APPLICATIONS

Not only become essential for basic genomic and molecular biology research

- Biotechnology and biomedical sciences

knowledge-based drug design, forensic DNA analysis, and agricultural biotechnology

- Computational studies of protein–ligand interactions provide a rational basis for the rapid identification of novel leads for synthetic drugs. Knowledge of the three-dimensional structures of proteins allows molecules to be designed that are capable of binding to the receptor site of a target protein with great affinity and specificity

- Forensics, results from molecular phylogenetic analysis have been accepted as evidence in criminal courts.

- Development of personalized and customized medicine.
- Doctor in a clinic to quickly sequence a patient’s genome and easily detect potential harmful mutations and to engage in early diagnosis and effective treatment of diseases
- Plant genome databases and gene expression profile analyses have played an important role in the development of new crop varieties that have higher productivity and more resistance to disease.

LIMITATIONS

It is no stretch in analogy that fighting diseases or other biological problems using bioinformatics is like fighting battles with intelligence

- Bioinformatics and experimental biology are independent, but complementary, activities. Bioinformatics depends on experimental science to produce raw data for analysis. It, in turn, provides useful interpretation of experimental data and important leads for further experimental research.
- Bioinformatics predictions are not formal proofs of any concepts. They do not replace the traditional experimental research methods of actually testing hypotheses.
- Quality of bioinformatics predictions depends on the quality of data and the sophistication of the algorithms

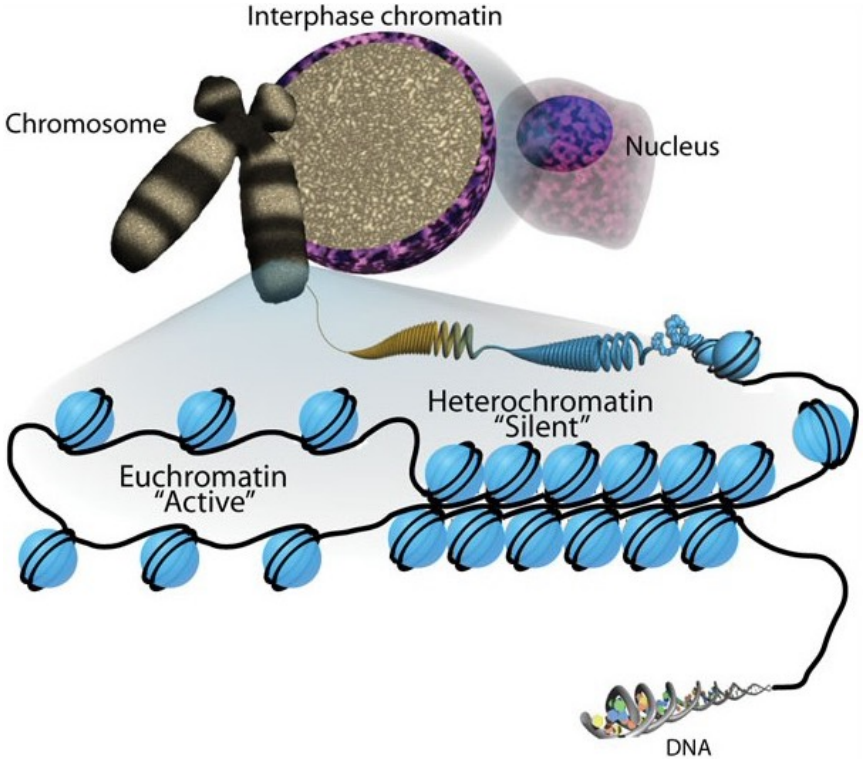
- Algorithms lack the capability and sophistication to truly reflect reality. They often make incorrect predictions that make no sense when placed in a biological context

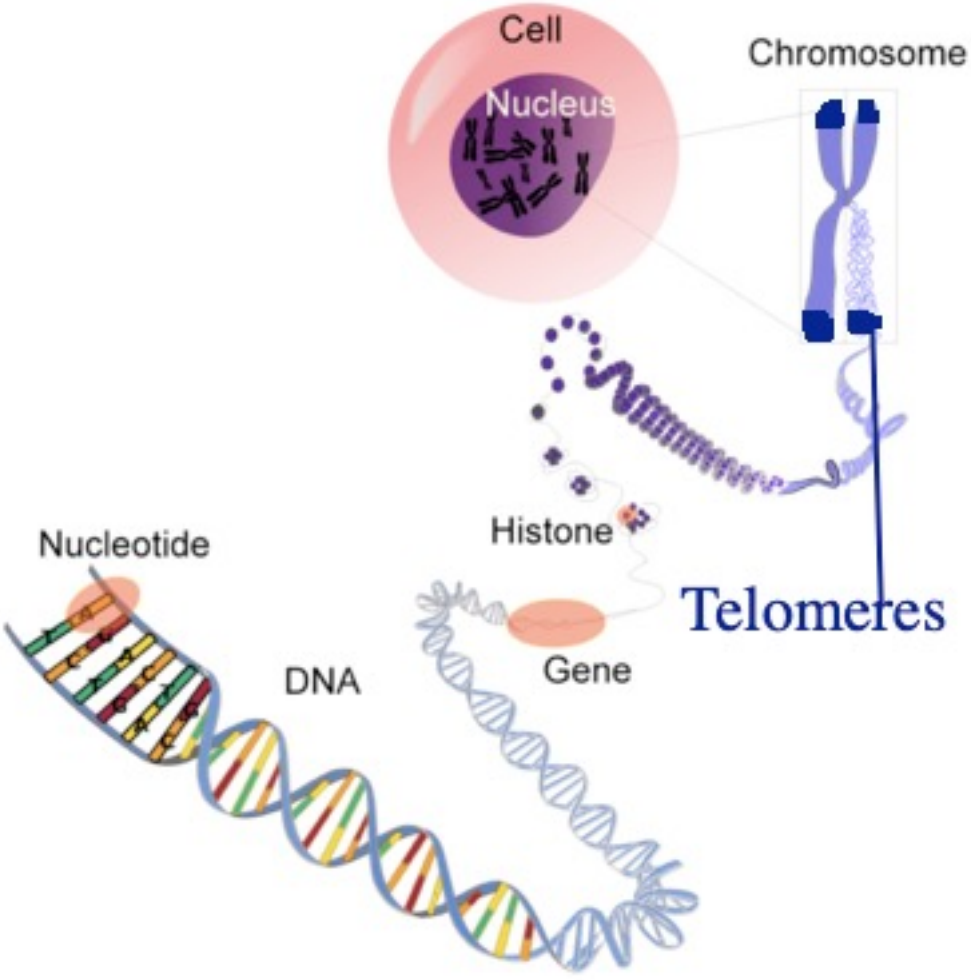
- The outcome of computation also depends on the computing power available

Many accurate but exhaustive algorithms cannot be used because of the slow rate of computation. Instead, less accurate but faster algorithms have to be used. This is a necessary trade-off between accuracy and computational feasibility

To gain a deeper understanding of cellular functions, mathematical models are needed to simulate a wide variety of intracellular reactions and interactions at the whole cell level. This molecular simulation of all the cellular processes is termed **Systems Biology.**

Introduction to Genome





DNA contains the genetic instructions specifying the development of **all cellular forms of life** and **most viruses**

Watson & Crick proposed the double helix structure of DNA in **1953**

Image source: Marjorie McCarty, Wikimedia Commons

DNA molecules consist of two chains (**strands**) of smaller molecules called **nucleotides**

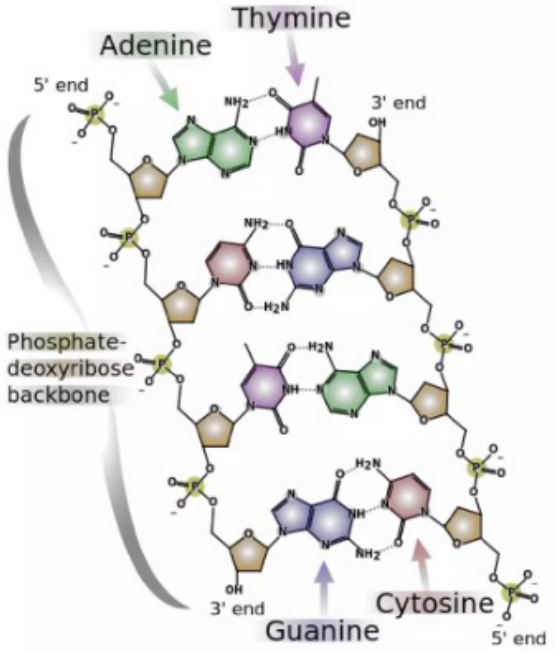


Image source:
Madeleine Price Ball,
Wikimedia
Commons

Each **nucleotide** consists of three parts: the sugar **deoxyribose**, a **phosphate** group, and one of four **bases**

The **bases** are **thymine T, adenine A, guanine G, cytosine C**

The sugars + phosphates form the backbone of the double helix

The four bases are molecules that contain rings which include both nitrogen (N) and carbon (C) atoms:

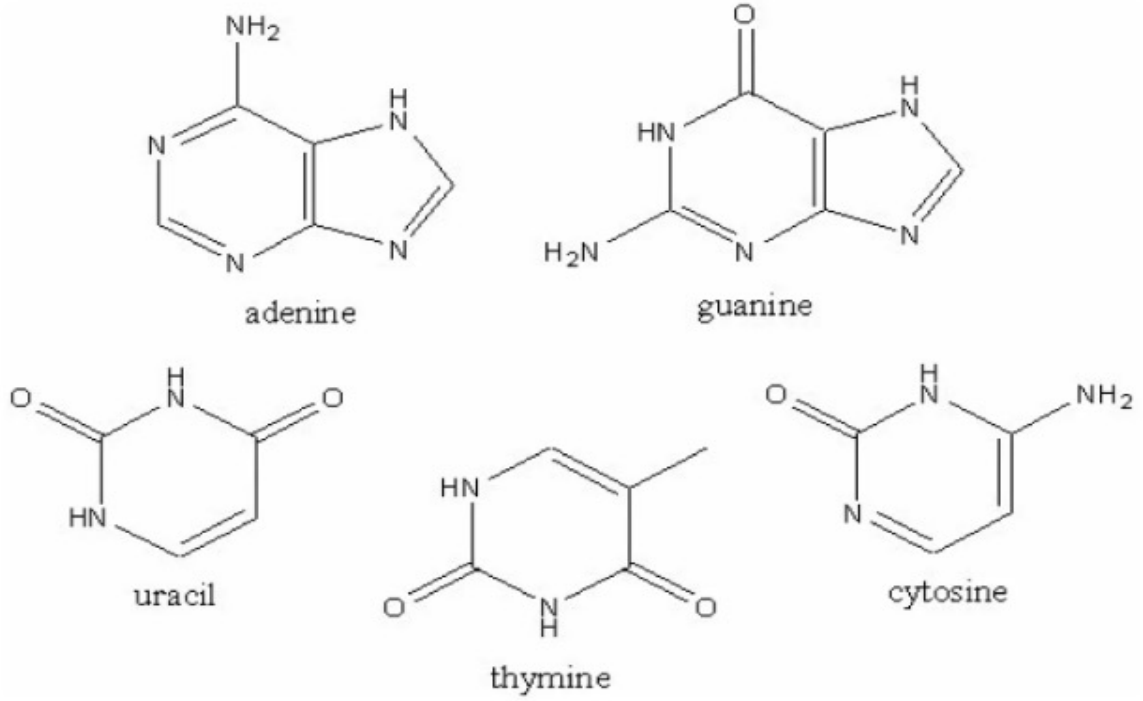


Image source: Mrbean427, Wikimedia Commons

The bases in the two strands of a DNA double helix are **complementary** to each other

T pairs with A, G pairs with C

Thus, if one strand has the sequence of bases TACG, the other strand must have the sequence of bases ATGC :

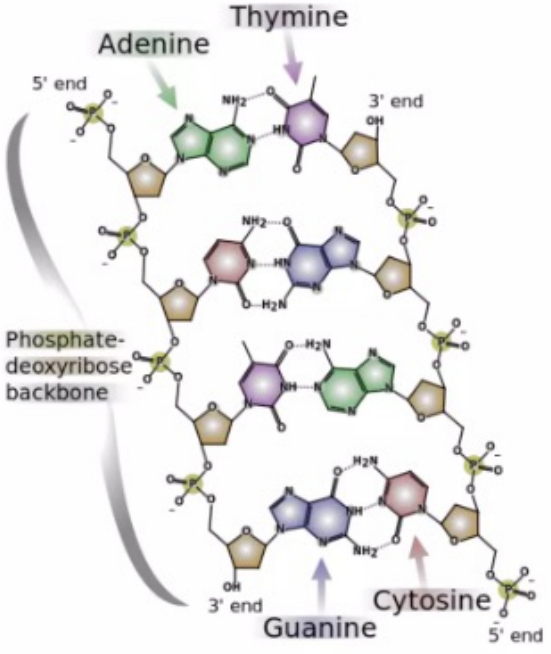


Image source:
Madeleine Price Ball,
Wikimedia
Commons

The 2 strands of DNA therefore contain **redundant information**

Each strand of DNA has **direction**

Each strand has **5' & 3' ends** (said "5-prime" and "3-prime")

The **5' end** is the end with a **terminal phosphate group**

In a DNA double helix, the 2 strands have opposite directions

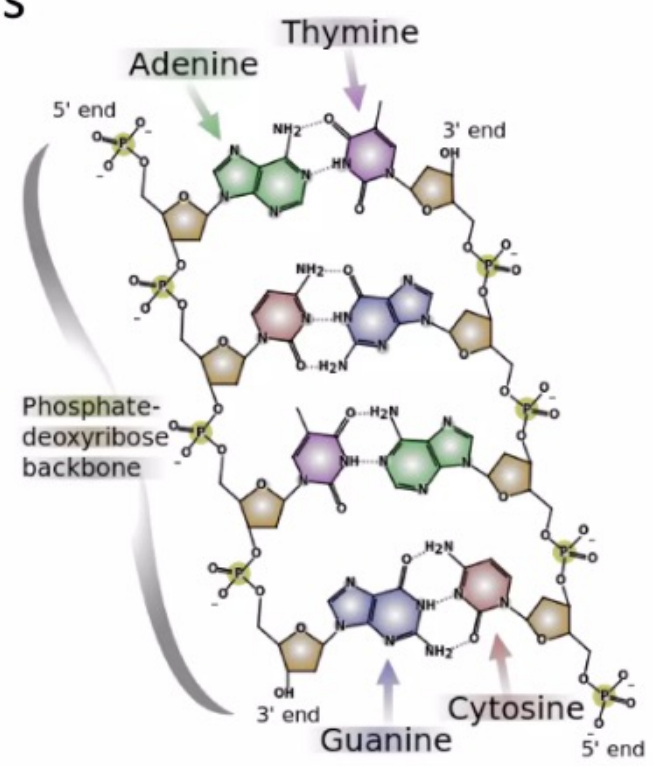


Image source:
Madeleine Price Ball,
Wikimedia
Commons

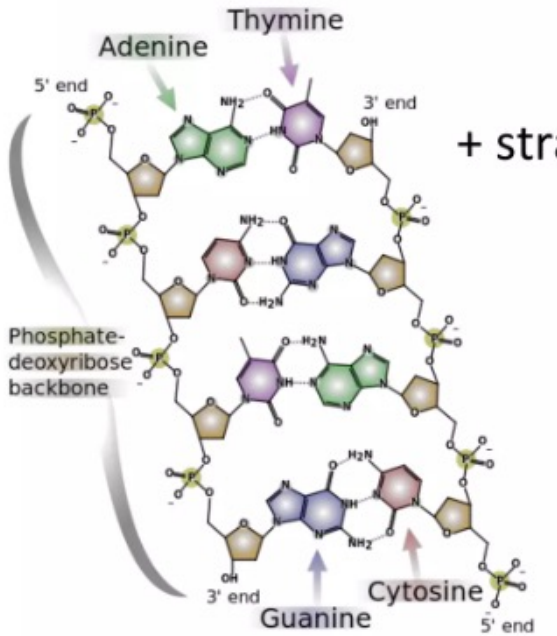
For convenience, one strand in a DNA double helix is called the **forward** or **+ (plus)** strand

Which strand to designate as '+' is decided by researchers studying the organism that the DNA is from

The choice is usually arbitrary, that is, there is no biological reason why one strand should be called the + strand

The other strand is called the **reverse** or **– (minus)** strand

- strand



+ strand

Image source: Madeleine Price Ball, Wikimedia Commons

By convention, we write a **DNA sequence** as the sequence of bases **from 5' to 3'**

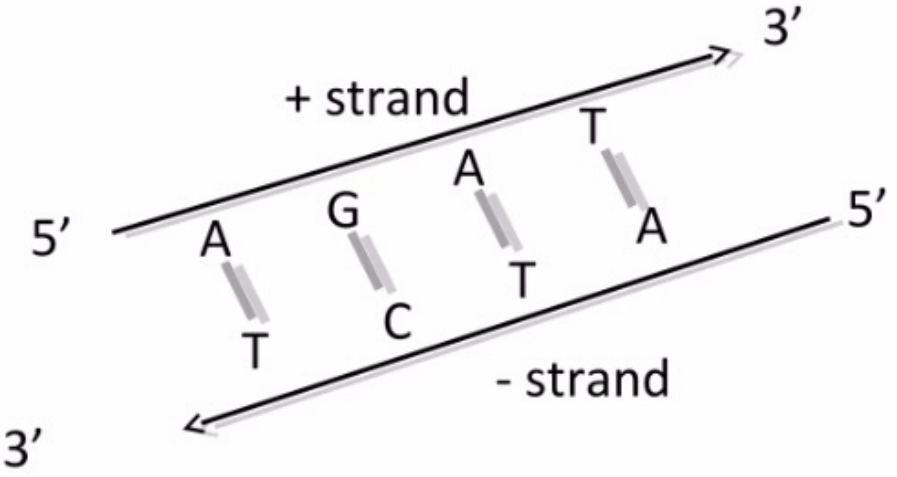
The sequence is for the + strand, unless otherwise specified

The – strand sequence can be inferred from the + strand sequence, as it's complementary to the + strand

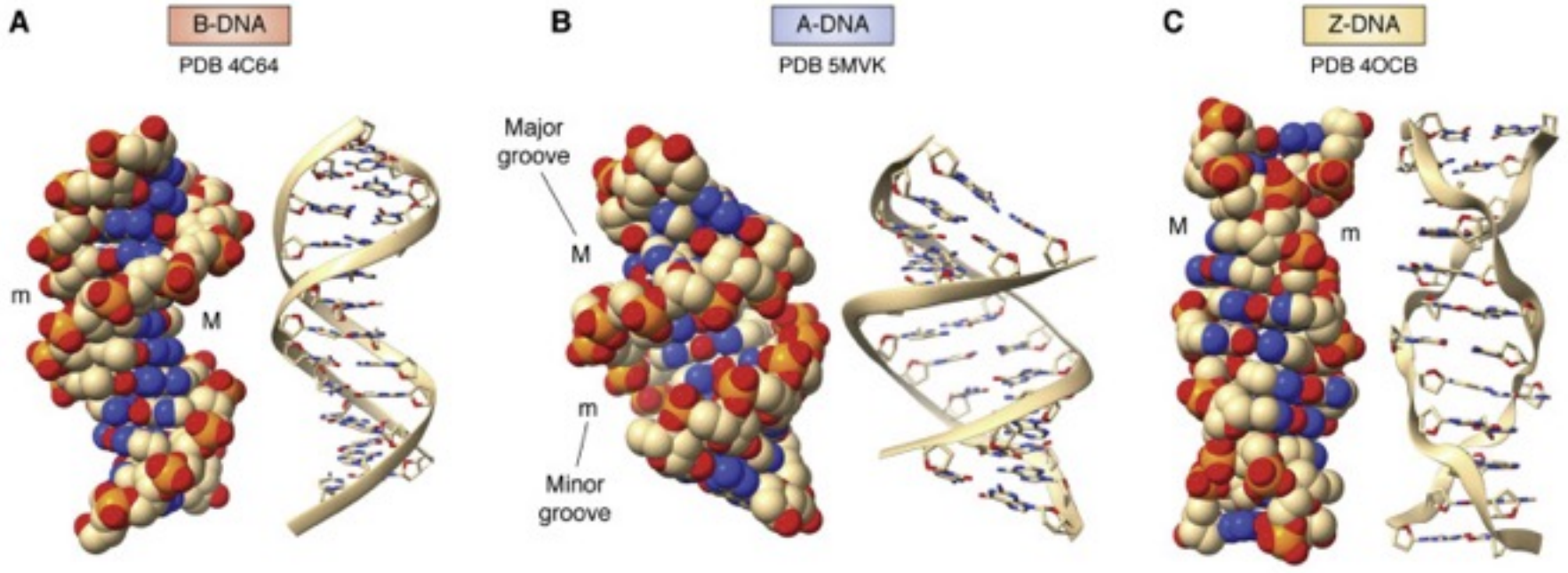
If the + strand sequence is 5'-AGAT-3', it's just written AGAT

The **– strand sequence** must be 3'-TCTA'-5 (the **complement**)

The – strand sequence 5'-ATCT'-3' is written ATCT (the **reverse complement**)



Types of DNA


A
B-DNA
PDB 4C64
m
M
B
A-DNA
PDB 5MVK
Major groove
M
m
Minor groove
C
Z-DNA
PDB 4OCB
M
m

- A **genome** is the set of all DNA in a cell

  A genome may consist of several **chromosomes**

  Each chromosome contains one long DNA molecule

  The DNA molecule in a chromosome can 1000s or millions of **base-pairs** long

  There are also many proteins bound to DNA, which act to **package the DNA** in a chromosome

- A chromosome is very tiny

  A chromosome that is **100 million base-pairs (bp) long is <0.01 mm**

  The human eye can only see objects of about **0.1 mm or larger**

One sesame seed: 2000-3000 µm (1 µm = 0.001 mm)
One grain of salt: 500 µm (0.5 mm)
Human egg cell: 130 µm (0.13 mm)
} **Visible with the human eye**

Human X chromosome: 7 µm (0.007 mm)
Size of one cell of the bacterium *Escherichia coli*: 3 x 0.6 µm
One 'A' (adenine): 0.0013 x 0.0008 µm
} **Invisible to the human eye**

See http://learn.genetics.utah.edu/content/begin/cells/scale/

- The human genome consists of **23 pairs of chromosomes**: 1-22, & XX (women)/XY (men)

The 23 chromosomes have **~3000 million base-pairs** of DNA

A cell has 46 chromosomes, so ~6000 million base-pairs

The largest is chromosome 1: 247 million bp (247 Mb)

The smallest is chromosome 22: 50 million bp (50 Mb)

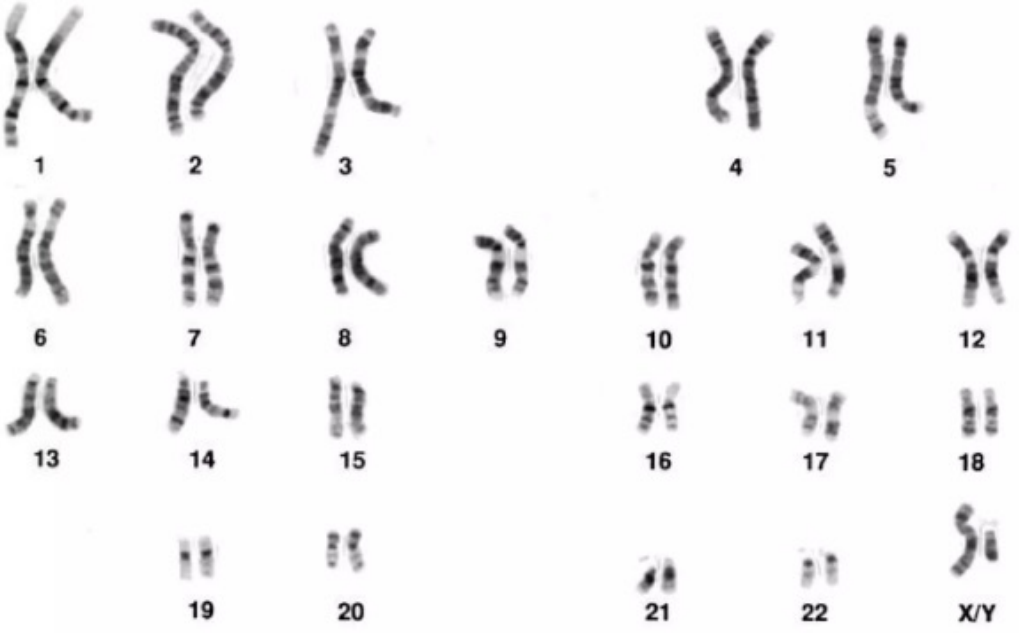


Image source: National Cancer Institute, Wikimedia Commons

- There is huge variation in chromosome number in the genomes of different species

  eg. the genome of the Australian ant *Myrmecia pilosula* consists of just **two pairs of chromosomes** (per cell)

- Some plants have a huge number of chromosomes

  eg. the genome of adder's tongue fern (*Ophioglossum reticulatum)* consists of **~720 pairs of chromosomes**

- Human chromosomes are **linear**, but many bacteria have 1 **circular** chromosome

  ie. the DNA molecule forms a large circle

  The bacterium *Escherichia coli* has a circular chromosome of **~5 million base-pairs** (5 Mb)

  Some bacteria have linear chromosomes eg. the bacterium ***Borrelia burgdorferi*** (which causes Lyme disease) has **one linear chromosome**

  Also, some bacteria have >1 chromosome eg. ***Rhodobacter sphaeroides*** has **2 circular chromosomes**

- As well as chromosomes, many bacteria have ≥1 small circular DNA molecules: **plasmids**

The bacterial **chromosome** is large (~0.5-13 Mb), & contains essential genes controlling cell development & structure

Plasmids are smaller (~0.1-0.5 Mb), and are usually not essential for the bacterium to survive

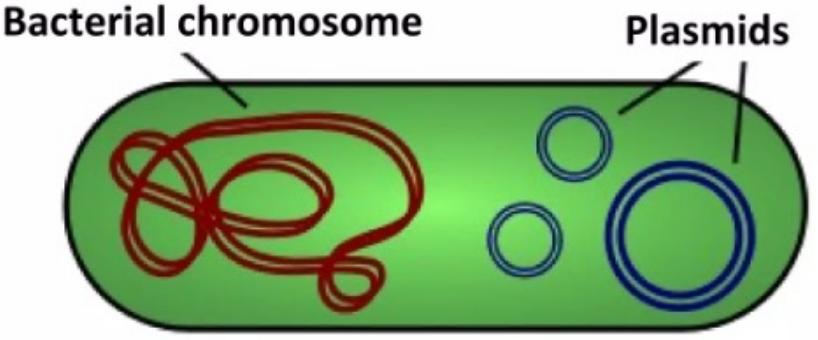


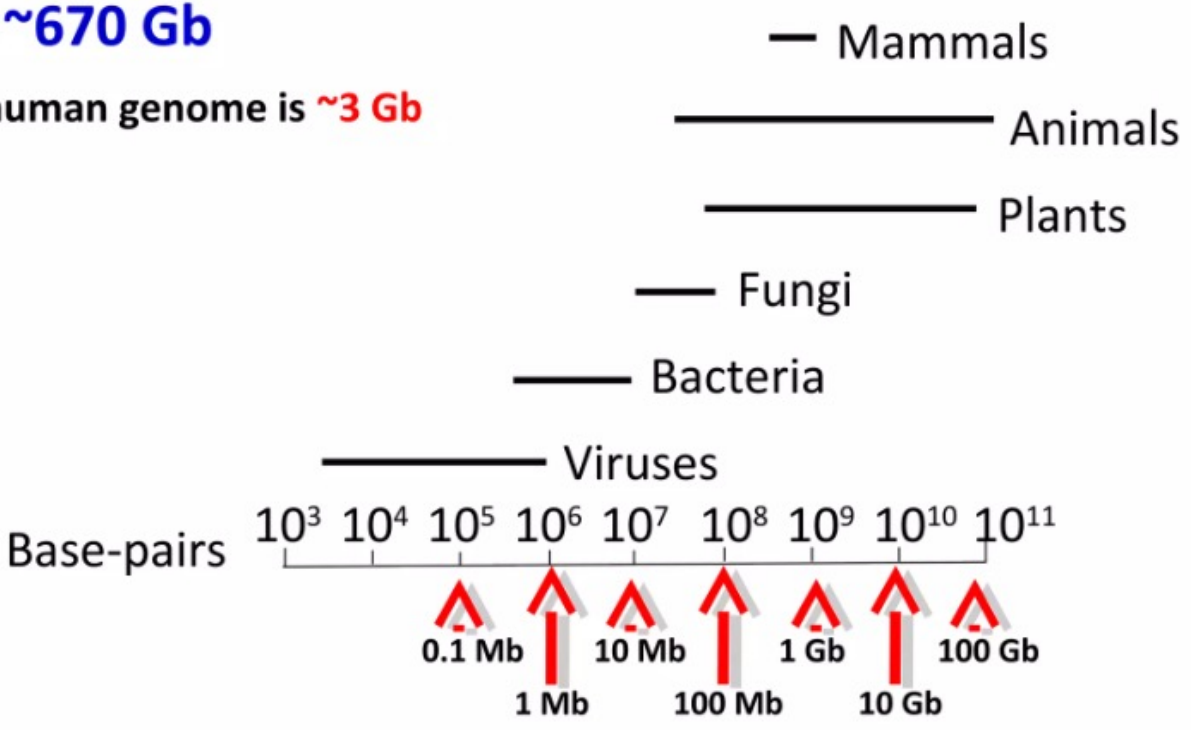
Image source: User:Spaully, Wikimedia Commons

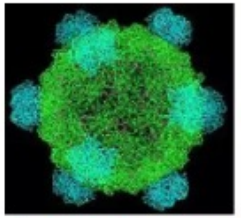
- Genome sizes are measured in **base-pairs (bp)**

  1 Mb (Megabase) = 1 million bp; **1 Gb** (Gigabase) = 1000 Mb

- Bacteria usually have 1 circular chromosome of **~0.5-13 Mb**
- Animals & plants & fungi have larger genomes, of **~8 Mb to ~670 Gb**

  e.g. **the human genome is ~3 Gb**





The virus
phiX174

Genome sequencing

Image source: Fdardel, Wikimedia Commons

- **DNA sequencing** means finding out the sequence of base-pairs along the double helix
- **Fred Sanger** received the Nobel Prize in **1980** for developing a method to sequence DNA

  Known as the **dideoxy method** or **Sanger method**
  Sanger also received a Nobel Prize (‘58) for sequencing proteins
- The first genomes sequenced were viruses
- Fred Sanger’s group in Cambridge sequenced the first virus in **1977**:

  Phage *phiX174*, has a 5386 base genome

Gene Prediction

- The ultimate goal is to describe all the genes computationally with near 100% accuracy.

- The ability to accurately predict genes can significantly reduce the amount of experimental verification work required.

- Gene prediction, in fact, represents one of the most difficult problems in the field of pattern recognition. This is because coding regions normally do not have conserved motifs. Detecting coding potential of a genomic region has to rely on subtle features associated with genes that may be very difficult to detect.

Ab-initio Gene Prediction

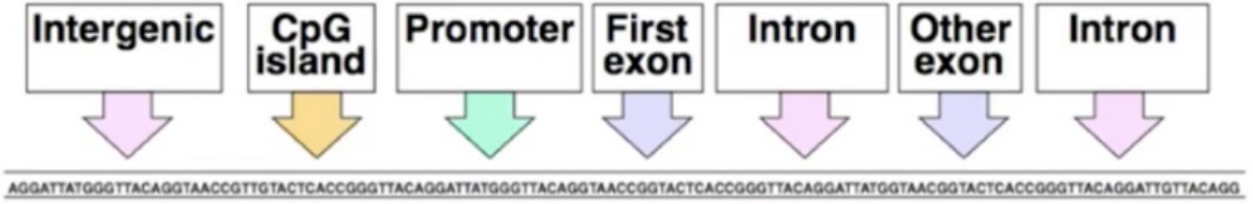
The ab-initio based approach predicts genes based on the given sequence alone. It does so by relying on two major features associated with genes.

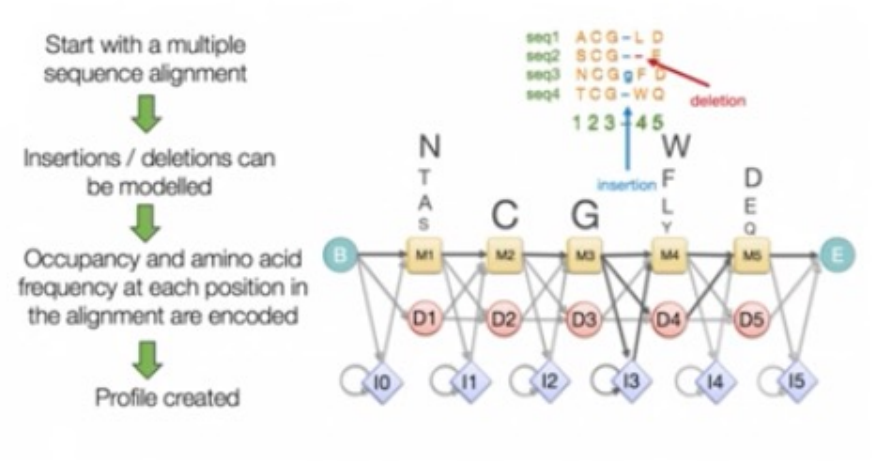
The first is the existence of gene signals, which include start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites, and polyadenylation (poly-A) sites.

In addition, the triplet codon structure limits the coding frame length to multiples of three, which can be used as a condition for gene prediction.

The second feature used by ab-initio algorithms is gene content, which is statistical description of coding regions. It has been observed that nucleotide composition and statistical patterns of the coding regions tend to vary significantly from those of the noncoding regions. The unique features can be detected by employing probabilistic models such as Markov models or hidden Markov models to help distinguish coding from noncoding regions.

HMM





- Gene prediction
- Pairwise & multiple sequence alignment
- Building a profile for a sequence family
- Base-calling
- Modeling DNA sequencing errors
- Protein secondary structure prediction
- Identifying copy number variations

SCORING MATRICES

In the dynamic programming algorithm presented, the alignment procedure has to make use of a scoring system, which is a set of values for quantifying the likelihood of one residue being substituted by another in an alignment. The scoring systems is called a *substitution matrix* and is derived from statistical analysis of residue substitution data from sets of reliable alignments of highly related sequences.

Scoring matrices for nucleotide sequences

A positive value or high score is given for a match and a negative value or low score for a mismatch. This assignment is based on the assumption that the frequencies of mutation are equal for all bases

substitutions between purines and purines or between pyrimidines and pyrimidines occur more frequently than transversions.

Therefore, a more sophisticated statistical model with different probability values to reflect the two types of mutations is needed

Scoring matrices for amino acid sequences

Scoring has to reflect the physicochemical properties of amino acid residues, as well as the likelihood of certain residues being substituted among true homologous sequences.

Certain amino acids with similar physicochemical properties can be more easily substituted than those without similar characteristics.

Substitutions among similar residues are likely to preserve the essential functional and structural features. However, substitutions between residues of different physicochemical properties are more likely to cause disruptions to the structure and function.

This type of disruptive substitution is less likely to be selected in evolution because it renders nonfunctional proteins.

- Phenylalanine, tyrosine, and tryptophan all share aromatic ring structures
- Arginine, lysine, and histidine
- Aspartic acid, glutamic acid, asparagine, and glutamine belong to the acid and acid amide groups
- Methionine, isoleucine, leucine, and valine
- Small and polar residues include serine, threonine, and cysteine

Cysteine contains a sulfhydryl group that plays a role in metal binding, active site, and disulfide bond formation. Substitution of cysteine with other residues therefore often abolishes the enzymatic activity or destabilizes the protein structure

Small and nonpolar residues such as glycine and proline are also unique in that their presence often disrupts regular protein secondary structures

Substitutions with these residues do not frequently occur .

TABLE 12.1. Twenty Standard Amino Acids Grouped by Their Common Side-Chain Features

Amino Acid Group	Amino Acid Name	Three- and One-Letter Code	Main Functional Features
Small and nonpolar	Glycine Alanine Proline	Gly, G Ala, A Pro, P	Nonreactive in chemical reactions; Pro and Gly disrupt regular secondary structures
Small and polar	Cysteine Serine Threonine	Cys, C Ser, S Thr, T	Serving as posttranslational modification sites and participating in active sites of enzymes or binding metal
Large and polar	Glutamine Asparagine	Gln, Q Asn, N	Participating in hydrogen bonding or in enzyme active sites
Large and polar (basic)	Arginine Lysine Histidine	Arg, R Lys, K His, H	Found in the surface of globular proteins providing salt bridges; His participates in enzyme catalysis or metal binding
Large and polar (acidic)	Glutamate Aspartate	Glu, E Asp, D	Found in the surface of globular proteins providing salt bridges
Large and nonpolar (aliphatic)	Isoleucine Leucine Methionine Valine	Ile, I Leu, L Met, M Val, V	Nonreactive in chemical reactions; participating in hydrophobic interactions
Large and nonpolar (aromatic)	Phenylalanine Tyrosine Tryptophan	Phe, F Tyr, Y Trp, W	Providing sites for aromatic packing interactions; Tyr and Trp are weakly polar and can serve as sites for phosphorylation and hydrogen bonding

Scoring matrices for amino acids

Amino acid substitution matrices, which are 20 × 20 matrices, have been devised to reflect the likelihood of residue substitutions

There are essentially two types of amino acid substitution matrices. One type is based on interchangeability of the genetic code or amino acid properties, and the other is derived from empirical studies of amino acid substitutions.

The empirical matrices, which include PAM and BLOSUM matrices, are derived from actual alignments of highly similar sequences. By analyzing the probabilities of amino acid substitutions in these alignments, a scoring system can be developed by giving a high score for a more likely substitution and a low score for a rare substitution.

For a given substitution matrix, a positive score means that the frequency of amino acid substitutions found in a data set of homologous sequences is greater than would have occurred by random chance. They represent substitutions of very similar residues or identical residues.

A zero score means that the frequency of amino acid substitutions found in the homologous sequence data set is equal to that expected by chance. In this case, the relationship between the amino acids is weakly similar at best in terms of physicochemical properties.

A negative score means that the frequency of amino acid substitutions found in the homologous sequence data set is less than would have occurred by random chance. This normally occurs with substitutions between dissimilar residues.

The substitution matrices apply logarithmic conversions to describe the probability of amino acid substitutions. The converted values are the so-called log-odds scores (or log-odds ratios), which are logarithmic ratios of the observed mutation frequency divided by the probability of substitution expected by random chance. The conversion can be either to the base of 10 or to the base of 2.

For example, in an alignment that involves ten sequences, each having only one aligned position, nine of the sequences are F (phenylalanine) and the remaining one I (isoleucine). The observed frequency of I being substituted by F is one in ten (0.1), whereas the probability of I being substituted by F by random chance is one in twenty (0.05). Thus, the ratio of the two probabilities is 2 (0.1/0.05). After taking this ratio to the logarithm to the base of 2, this makes the log odds equal to 1. This value can then be interpreted as the likelihood of substitution between the two residues being 2 1 , which is two times more frequently than by random chance.

PAM : Point Accepted Mutation

- **Margarett Dayhoff (1978)**
- Based on evolutionary distance obtained from 71 closely related protein sequence alignments
- Muataion that comprise of change in single amino acid (substitution) which is accepted by natural selection. (Accepted point mutation)
  - ✓ Mutation of gene region (coding single amino acid) to produce different amino acid.
  - ✓ That mutation accepted as predominant form in a species.
- 1 PAM meaning **one APM per 100** amino acids.
- It is based on global alignment (aligns entire sequence)

PAM : Point Accepted Mutation

- Markovian assumption that each amino acid change at a site being independent of previous change at that site.
- So we can cover as much as evolutionary divergence as we need (higher PAM unit) by extrapolating same PAM1 again and again.

- 1 PAM denoted as PAM_1
- $PAM_1 \times PAM_1 = PAM_2$

- So generally,

$PAM_x = PAM_1^x$ (x iteration of PAM_1)

$PAM_{250} = PAM_1^{250}$ (widely used scoring matrices)

PAM : Point Accepted Mutation

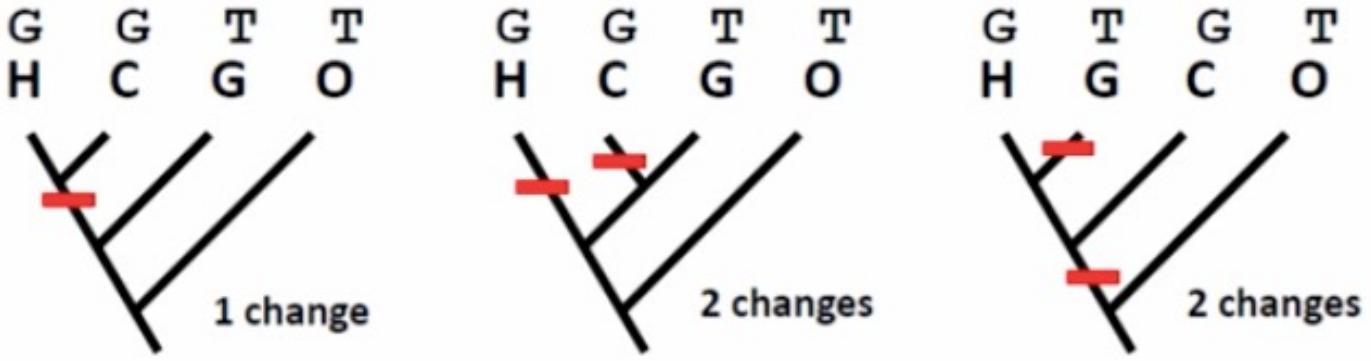
If we consider PAM_{100}, it does not mean that after 100 PAM of evolution every residue will have change.

- ✓ Some may mutate several times.
- ✓ Some may returned to its original state.
- ✓ Some residue may not changed at all.

PAM matrix origin

- Based on 71 groups of closely related protein.
- PAM (percent accepted mutation) is inferred from the types of changes observed in this proteins. (tabulated)
- Relative mutability of different amino acids were calculated.
- These two data combined in mutation probability matrix.
- The elements of this matrix give the probability that the amino acid in one column will be replaced by the amino acid in some row after a given evolutionary interval.
- 0 PAM having ‘ones’ on the main diagonal and ‘zeroes’ elsewhere.

Parsimony: an example



Globin pseudogene sequences:

informative site

Human:	CACAATA...TGAGC..GAAGAGATG...GTGAAG
Chimp:	CACAATA...GGAGC..GAAGAGACG...GTGAAG
Gorilla:	CACAATA...TGAGT..TAAGAGACG...TTGAAG
Orang:	CACAATA...TGAGT..TAAGAGACA...TTGAAT

TABLE 3.1. Correspondence of PAM Numbers with Observed Amino Acid Mutational Rates

PAM Number	Observed Mutation Rate (%)	Sequence Identity (%)
0	0	100
1	1	99
30	25	75
80	50	50
110	40	60
200	75	25
250	80	20

C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-2	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	**S**	**T**	**P**	**A**	**G**	**N**	**D**	**E**	**Q**	**H**	**R**	**K**	**M**	**I**	**L**	**V**	**F**	**Y**	**W**

Figure 3.5: PAM250 amino acid substitution matrix. Residues are grouped according to physicochemical similarities.

-It is *based on PROSITE signatures* (signatures are short expressions like C-X-X-C-X-X-X-C). In short BLOSUM approach is as follows-

Series of blocks amino acid substitution matrices are derived *based on the direct observation for every possible amino acid substitution in multiple sequence alignments.*

- These were constructed *based on more than 2000 conserved amino acid patterns* (locally aligned each feature to get 'blocks') *representing 500 groups* of protein sequences.

- Blocks are *locally conserved* regions/ *ungapped alignments of less than sixty* amino acid residues.

- More constrained regions are likely to be *related to structure/function*.

- Blocks contain sequences *at all different evolutionary distances* and may be highly biased (e.g. many identical sequences)

- The *frequencies of amino acid substitutions of residues* in these blocks are calculated to *produce a numerical table, or block* substitution matrix. It deals with bias and distance.

Algorithm is as follows-

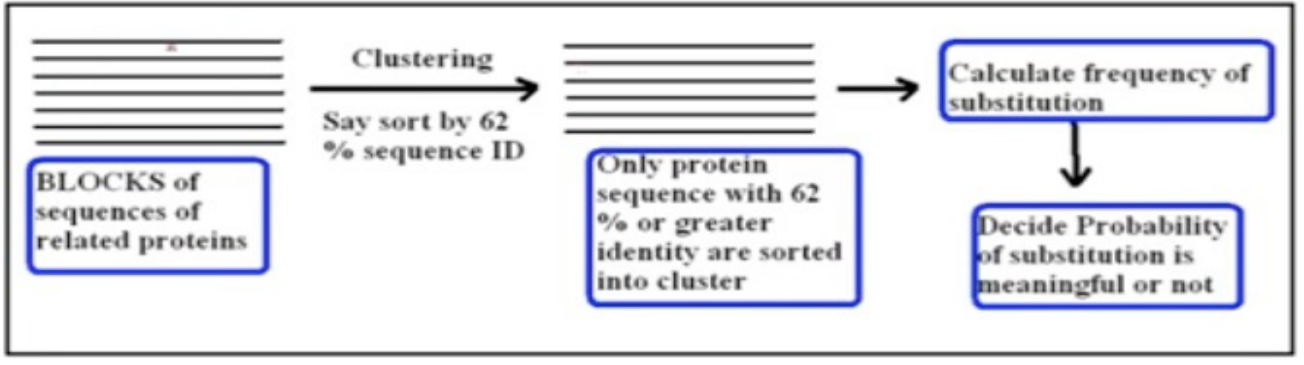
- *Cluster* all sequences with less than X% identities.

- Clustered sequences *count as 1 sequence*.

- If *X is 100%* it simply *removes identical sequences* If X *is <100%* it reduces the *weight on closely related sequences*.

- *Calculate substitution frequencies and log-odd matrix*. This gives a BLOSUM X table.

- A clustering approach sorts the sequences in each block to closely related groups
- The frequency of substitution of residues within these families derives the probability of meaningful substitution as shown in following figure:



Sequence	Position			
Seq1	B	A	B	A
Seq2	A	A	A	C
Seq3	A	A	C	C
Seq4	A	A	B	A
Seq5	A	A	C	C
Seq6	A	A	B	C

Step1: count the frequency of each amino acid

AA	Frequency of occurrence
A	14
B	4
C	6

Total = 24

Step2: count the frequency of each amino acid pair

Pair	Frequency of occurrence
AA	26
AB	8
AC	10
BB	3
BC	6
CC	7

Total = 60

Step3: count the observed frequency of amino acid pairs

$AB_{obs} = 8/60$

Step4: count the expected frequency of amino acid pairs

$AB_{exp} = (14/24 \times 4/24) \times 2$

Step5: calculate the log-odd ratio

$2\log_2(AB) = 2 \times (8/60)/(196/576)$

Pair	Observed (O)	Expected (E)	$2\log_2(O/E)$
AA	26/60	196/576	0.70
AB	8/60	112/576	-1.09
AC	10/60	168/576	-1.61
BB	3/60	16/576	1.70
BC	6/60	48/576	0.53
CC	7/60	36/576	1.80

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

Figure 3.6: BLOSUM62 amino acid substitution matrix.

- The BLOSUM matrices are *actual percentage identity values of sequences selected for construction of matrices.*

- In the *reversing order of the PAM numbering* system, *the lower the BLOSUM number*, the *more divergent* sequences they represent.

- For example in *BLOSUM 62-sequences greater than 62% identical are clustered* and in BLOSUM 80-sequences greater than 80% identical are clustered

A MULTIPLE SEQUENCE ALIGNMENT OF 5 AMINO ACID SEQUENCES

HIGHLY CONSERVED RESIDUES

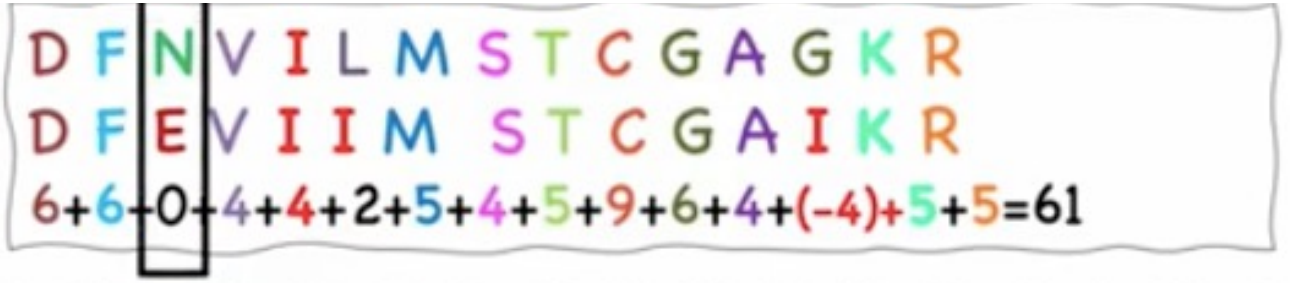
Percentage identity among sequences decides the Matrix type
For example, **BLOSUM-62 has been derived from amino acid sequences with 62% identity**

- The **numerical value** (ex 45 or 62) associated with BLOSUM matrix **represents the cutoff** of the clustering step
- The value of 62 indicates that sequences were put into the same cluster if they were more than 62% identical.
- If one wants **more diverse sequences to be included** in the cluster, **lower cut off values must be selected**, because lower cutoff values represents longer evolutionary timescales.
- Hence matrices with lower cutoff values are appropriate for **seeking more distant relationships**
- BLOSUM 62 is standard matrix for ungapped alignments, while BLOSUM 50 is more commonly used when generating gapped alignments.

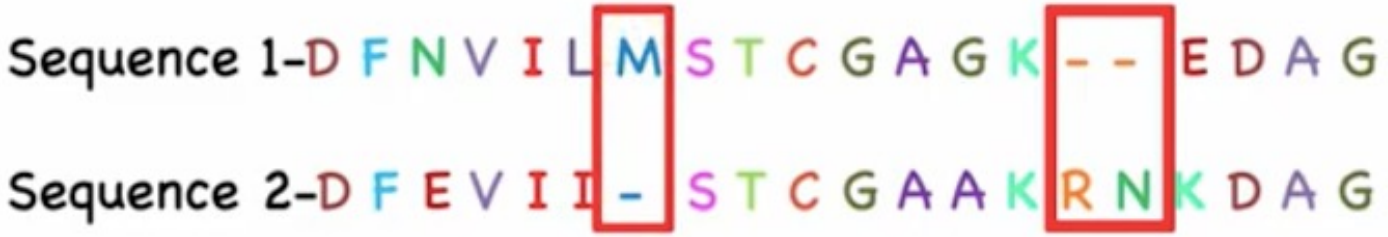
- The BLOSUM score------- for a *particular residue pair* is *derived from* the *log ratio of observed residue substitution frequency versus the expected probability of a particular residue.*

- The *log odds is taken* to the *base of 2* (instead of 10 as in the PAM matrices). The *resulting value is rounded to the nearest integer* and entered into the substitution matrix.

- *Positive score* corresponds to *substitutions that occur more frequently than expected among evolutionarily conserved replacements* and *reverse* is true for *negative scores*.



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



Gap score for the first gap, n=1, A= 11, B=1
=A + Bn
= 11 + 1x1 = 12

Gap score for the second gap, here, n=2
=A + Bn
= 11 + 1x2 = 13

Gap existence → introduction of a gap.
a given fixed value "A" is deducted
For PROTEIN default "A" value in BLAST-p → 11

Gap extension → how big is the gap.
A given fixed value "B" x number of gaps "n"
For PROTEIN default "B" value in BLAST-p → 1

Gap score = A + Bn

Sequence 1-D F N V I L M S T C G A G K - - E D A G

Alignment- D F V I + S T C G A K D A G

Sequence2- D F E V I I - S T C G A A K R N K D A G

6+6+0+4+4+2-12+4+5+9+6+4+(-4)+5-13+1+6+4+6=43

Matches/Identities→a letter represents an identical match

Conservative Substitution→ a + sign

Gaps and scores with "0" or a negative value are not represented and are left blank.

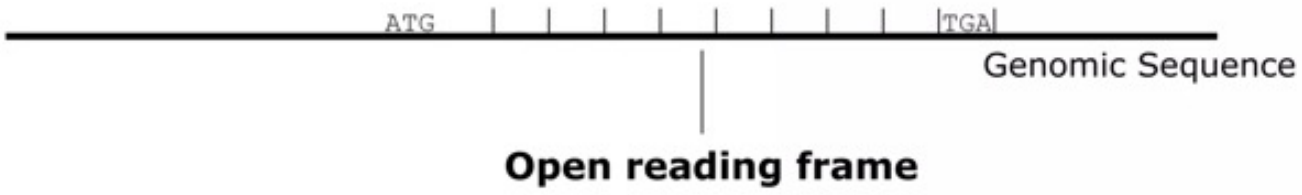
Max score = Σ(identities, substitutions) – Σ (gap penalties)

= [69 + {0+2+(-4)+1}] – (12+13)

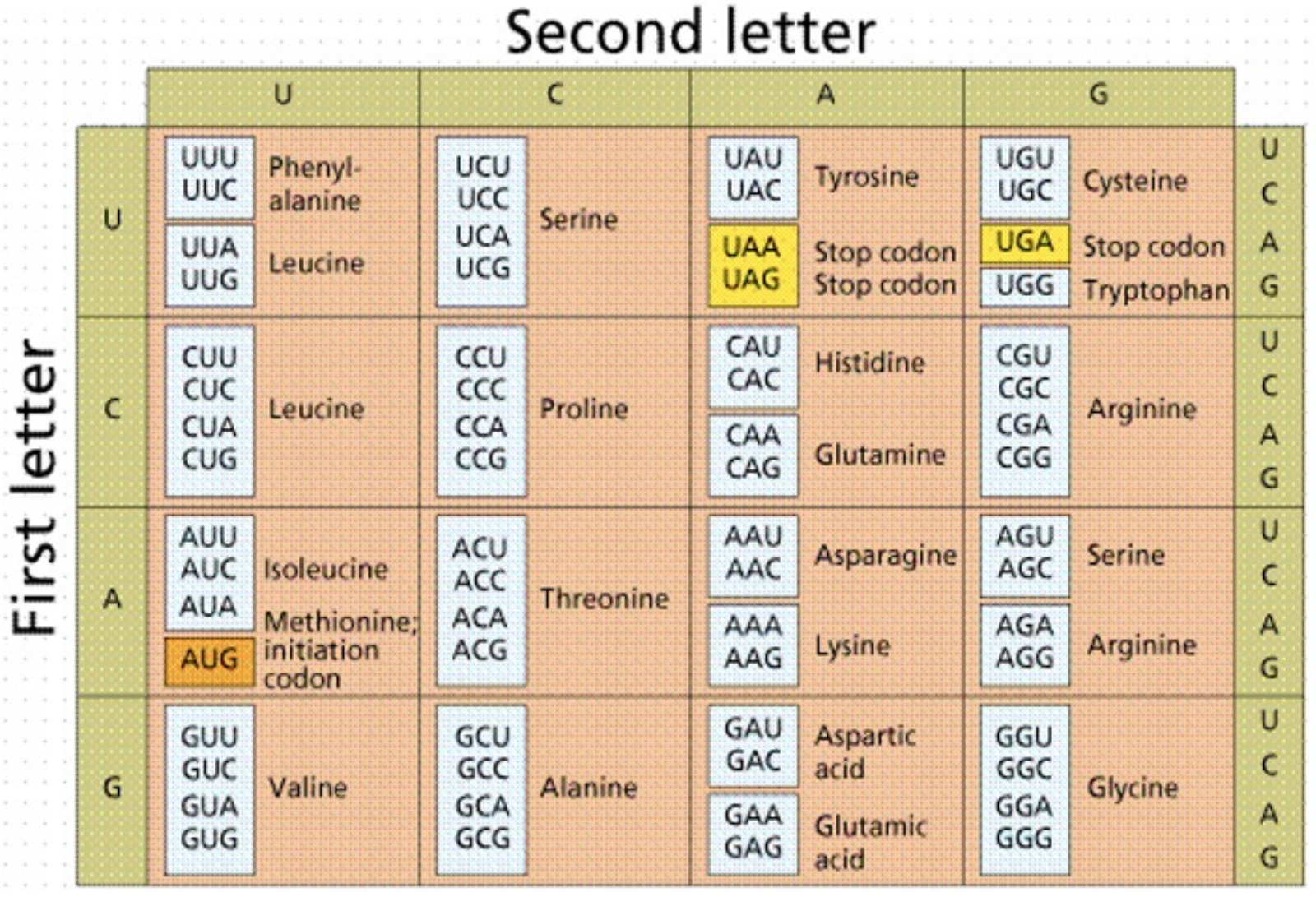
= 68 – 25 = 43

ORFs

- In molecular genetics, an **open reading frame** (ORF) is the part of a reading frame that contains no stop codons or region of amino acids coding triplet codons
- **An ORF starts with a start codon ATG (Met) in most species and ends with a stop codon (TAA, TAG, TGA)**
- Transcription termination pause site is located after ORF



- One common use of open reading frames is as one piece of evidence to assist in gene prediction
- Potential coding regions of a gene are detected by looking at **ORFs** in a DNA sequence


Second letter
First letter
U
C
A
G
UUU UUC Phenyl-alanine
UUA UUG Leucine
UCU UCC UCA UCG Serine
UAU UAC Tyrosine
UAA UAG Stop codon Stop codon
UGU UGC Cysteine
UGA Stop codon
UGG Tryptophan
CUU CUC CUA CUG Leucine
CCU CCC CCA CCG Proline
CAU CAC Histidine
CAA CAG Glutamine
CGU CGC CGA CGG Arginine
AUU AUC AUA Isoleucine
AUG Methionine; initiation codon
ACU ACC ACA ACG Threonine
AAU AAC Asparagine
AAA AAG Lysine
AGU AGC Serine
AGA AGG Arginine
GUU GUC GUA GUG Valine
GCU GCC GCA GCG Alanine
GAU GAC Aspartic acid
GAA GAG Glutamic acid
GGU GGC GGA GGG Glycine

Identifying ORFs (Open Reading Frames)

Simple First Step In Gene Finding

- A genome of length *n* is comprised of (*n*/3) codons
- Stop codons break genome into segments between consecutive Stop codons
- Segments of these that start from Start codon (**ATG**) are ORFs
  - ORFs in different frames may overlap

- Genomic sequence is translated in codon frames
- Stop codons are identified in each frame
- Regions without stop codons are called ORFs
- All of the likely ORFs in a sequence are located and tagged
- Longest ORF from a Met codon is a good prediction of a protein encoding sequence

Reading sequence of DNA in six reading frames

- Every region of DNA has six possible **reading frames;**
  - Three in the forward direction
  - Three in the reverse direction

Frame **1** starts with the "**a**", Frame **2** with the "**t**" and Frame **3** with the "**g**"

5' atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa 3'

1 atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
M P K L N S V E G F S S F E D D V *

2 tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat
C P S * I A * R G F H H L R T M Y

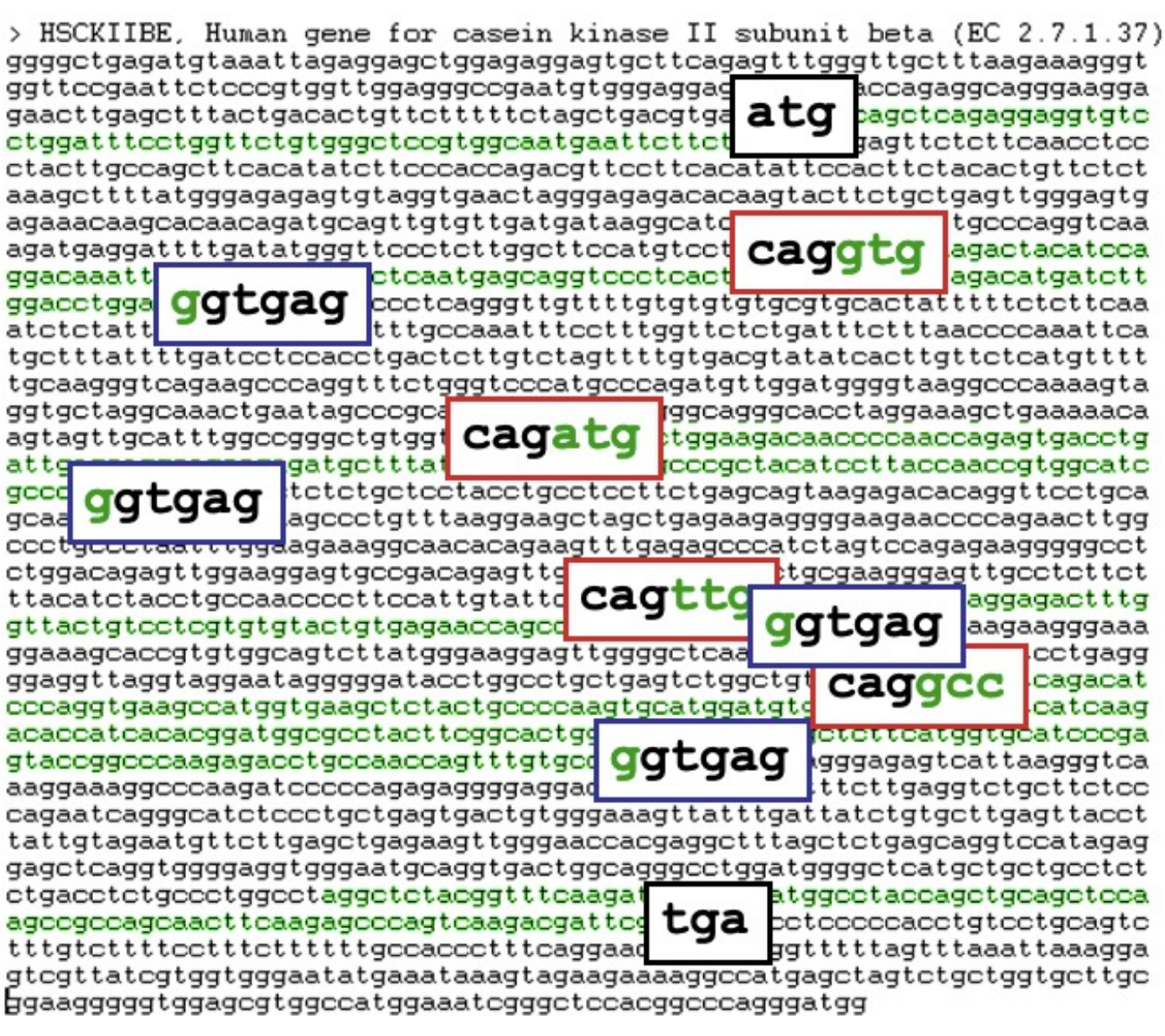
3 gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata
A Q A E * R R G V F I I * G R C I

Stop codons are indicated by an "*" in the protein sequence

The longest ORF is in Frame **1**

> HSCKIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgcttcagagtttgggttgctttaagaaagggt
ggttccgaattctcccgtggttggagggccgaatgtgggaggagggaggataccagaggcagggaagga
gaacttgagctttactgacactgttctttttctagctgacgtgaagatgagcagctcagaggaggtgtc
ctggatttcctggttctgtgggctccgtggcaatgaattcttctgtgaagtgagttctcttcaacctcc
ctacttgccagcttcacatatcttcccaccagacgttccttcacatattccacttctacactgttctct
aaagcttttatgggagagagtgtaggtgaactagggagagacacaagtacttctgctgagttgggagtg
agaaacaagcacaacagatgcagttgtgttgatgataaggcatcacttagagcattttgcccaggtcaa
agatgaggattttgatatgggttccctcttggcttccatgtcctgacaggtggatgaagactacatcca
ggacaaatttaatcttactggactcaatgagcaggtccctcactatcgacaagctctagacatgatctt
ggacctggagcctggtgaggcaccctcagggttgttttgtgtgtgtgcgtgcactatttttctcttcaa
atctctattcacttgcctgaattttgccaaatttcctttggttctctgatttctttaaccccaaattca
tgctttattttgatcctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccaggtttctgggtcccatgcccagatgttggatggggtaaggcccaaaagta
ggtgctaggcaaactgaatagcccgcagcccctggatatgggcagggcacctaggaaagctgaaaaaca
agtagttgcatttggccgggctgtggttcagatgaagaactggaagacaaccccaaccagagtgacctg
attgagcaggcagccgagatgctttatggattgatccacgcccgctacatccttaccaaccgtggcatc
gcccagatggtgaggcctctctgctcctacctgcctccttctgagcagtaagagacacaggttcctgca
gcaagaagtcatgtttaagccctgtttaaggaagctagctgagaagaggggaagaaccccagaacttgg
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccatctagtccagagaaggggcct
ctggacagagttggaaggagtgccgacagagttggtatgggttgggctgcgaagggagttgcctcttct
ttacatctacctgccaacccccttccattgtattcacctcagttggaaaagtaccagcaaggagactttg
gttactgtcctcgtgtgtactgtgagaaccagccaatgcttcccattGgtgagtgttgaagaagggaaa
ggaaagcaccgtgtggcagtcttatgggaaggagttggggctcaacacattggagcctgagtcctgagg
ggaggttaggtaggaataggggataccttggcctgctgagtctggctgtctcccaggcctttcagacat
cccaggtgaagccatggtgaagctctactgccccaagtgcatggatgtgtacacacccaagtcatcaag
acaccatcacacggatggcgcctacttcggcactggtttccctcacatgctcttcatggtgcatcccga
gtaccggcccaagagacctgccaaccagtttgtgcccaggtagggagcagggagagtcattaagggtca
aaggaaaggcccaagatcccccagagaggggaggacagggcatggccctttcttgaggtctgcttctcc
cagaatcagggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagttgggaaccacgaggctttagctctgagcaggtccatagag
gagctcaggtggggaggtgggaatgcaggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaagatccatccgatggcctaccagctgcagctcca
agccgccagcaacttcaagagcccagtcaagacgattcgctgattccctcccccacctgtcctgcagtc
tttgtcttttccttttcttttttgccacccttcaggaaccctgtatggtttttagtttaaattaaagga
gtcgttatcgtggtgggaatatgaaataaagtagaagaaaaggccatgagctagtctgctggtgcttgc
tgaaggggtggagcgtggccatggaaatcgggctccacggcccagggatgg

> HSCKIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgcttcagagtttgggttgctttaagaaagggt
ggttccgaattctcccgtggttggagggccgaatgtgggaggagggaggataccagaggcagggaagga
gaacttgagctttactgacactgttctttttctagctgacgtgaagatgagcagctcagaggaggtgtc
ctggatttcctggttctgtgggctccgtggcaatgaattcttctgtgaagtgagttctcttcaacctcc
ctacttgccagcttcacatatcttcccaccagacgttccttcacatattccacttctacactgttctct
aaagcttttatgggagagagtgtaggtgaactagggagagacacaagtacttctgctgagttgggagtg
agaaacaagcacaacagatgcagttgtgttgatgataaggcatcacttagagcattttgcccaggtcaa
agatgaggattttgatatgggttccctcttggcttccatgtcctgacaggtggatgaagactacatcca
ggacaaatttaatcttactggactcaatgagcaggtccctcactatcgacaagctctagacatgatctt
ggacctggagcctggtgaggcaccctcagggttgttttgtgtgtgtgcgtgcactatttttctcttcaa
atctctattcacttgcctgaattttgccaaatttcctttggttctctgatttctttaaccccaaattca
tgctttattttgatcctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccaggtttctgggtcccatgcccagatgttggatggggtaaggcccaaaagta
ggtgctaggcaaactgaatagcccgcagcccctggatatgggcagggcacctaggaaagctgaaaaaca
agtagttgcatttggccgggctgtggttcagatgaagaactggaagacaaccccaaccagagtgacctg
attgagcaggcagccgagatgctttatggattgatccacgcccgctacatccttaccaaccgtggcatc
gcccagatggtgaggcctctctgctcctacctgcctccttctgagcagtaagagacacaggttcctgca
gcaagaagtcatgtttaagccctgtttaaggaagctagctgagaagaggggaagaaccccagaacttgg
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccatctagtccagagaaggggcct
ctggacagagttggaaggagtgccgacagagttggtatgggttgggctgcgaagggagttgcctcttct
ttacatctacctgccaacccccttccattgtattcacctcagttggaaaagtaccagcaaggagactttg
gttactgtcctcgtgtgtactgtgagaaccagccaatgcttcccattGgtgagtgttgaagaagggaaa
ggaaagcaccgtgtggcagtcttatgggaaggagttggggctcaacacattggagcctgagtcctgagg
ggaggttaggtaggaatagggggatacctggcctgctgagtctggctgtctcccaggcctttcagacat
cccaggtgaagccatggtgaagctctactgccccaagtgcatggatgtgtacacacccaagtcatcaag
acaccatcacacggatggcgcctacttcggcactggtttccctcacatgctcttcatggtgcatcccga
gtaccggcccaagagacctgccaaccagtttgtgcccaggtagggagcagggagagtcattaagggtca
aaggaaaggcccaagatcccccagagaggggaggacagggcatggcccttttcttgaggtctgcttctcc
cagaatcagggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagttgggaaccacgaggctttagctctgagcaggtccatagag
gagctcaggtggggaggtgggaatgcaggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaagatccatccgatggcctaccagctgcagctcca
agccgccagcaacttcaagagcccagtcaagacgattcgctgattccctcccccacctgtcctgcagtc
tttgtcttttccttttcttttttgccacccttcaggaaccctgtatggtttttagtttaaattaaagga
gtcgttatcgtggtgggaatatgaaataaagtagaagaaaaggccatgagctagtctgctggtgcttgc
ggaaggggtggagcgtggccatggaaatcgggctccacggcccagggatgg


> HSCKIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgcttcagagtttgggttgctttaagaaagggt
ggttccgaattctcccgtggttggagggccgaatgtgggaggag
gaacttgagctttactgacactgttctttttctagctgacgtga
ctggatttcctggttctgtgggctccgtggcaatgaattcttct
ctacttgccagcttcacatatcttcccaccagacgttccttcacatattccacttctacactgttctct
aaagcttttatgggagagagtgtaggtgaactagggagagacacaagtacttctgctgagttgggagtg
agatgaggattttgatatgggttccctcttggcttccatgtcct
ggacaaatt
ggacctgga
atctctatt
tgctttattttgatcctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccaggtttctgggtcccatgcccagatgttggatggggtaaggcccaaaagta
ggtgctaggcaaactgaatagcccgca
agtagttgcatttggccgggctgtggt
attg
gccc
gcaa
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccatctagtccagagaaggggcct
ctggacagagttggaaggagtgccgacagagttg
ttacatctacctgccaacccctccattgtattc
gttactgtcctcgtgtgtactgtgagaaccagcc
ggaaagcaccgtgtggcagtcttatgggaaggagttggggctcaa
ggaggttaggtaggaataggggatacctggcctgctgagtctggctgt
cccaggtgaagccatggtgaagctctactgccccaagtgcatggatgtg
acaccatcacacggatggcgcctacttcggcactgg
gtaccggcccaagagacctgccaaccagtttgtgcc
aaggaaaggcccaagatcccccagagaggggaggac
cagaatcagggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagttgggaaccacgaggctttagctctgagcaggtccatagag
gagctcaggtggggaggtgggaatgcaggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaaga
agccgccagcaacttcaagagcccagtcaagacgattcg
tttgtcttttccttctttttgccacccttcaggaac
gtcgttatcgtggtgggaatatgaaataaagtagaagaaaaggccatgagctagtctgctggtgcttgc
ggaaggggtggagcgtggccatggaaatcgggctccacggcccagggatgg
atg
caggtg
ggtgag
cagatg
ggtgag
cagttg
ggtgag
caggcc
ggtgag
tga

SOFTWARES TO FIND ORFs

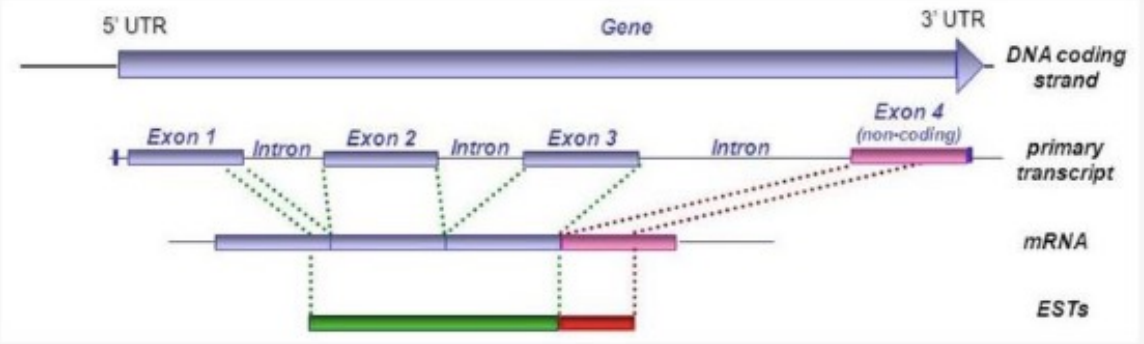
- ORF Finder: It is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database
- ORF Investigator: It is a program which not only gives information about the coding and non-coding sequences but also can perform pairwise global alignment of different gene/DNA regions sequences
- ORF Predictor: It is a web server designed for identifying protein-coding regions in expressed sequence tag (EST)-derived sequences

BACKGROUND

- Late 80's: GenBank estimates it will take 10-12 years to sequence the ~3 billion *bp* human genome.
- Roughly 3% of the DNA are coding sequences for genes.
- Complete DNA sequencing is expensive and time consuming.
- 1983: Putney et al. first used cDNA sequencing in genome identification
- 1991: Adams et al. found 230 ESTs, representing new neurological genes

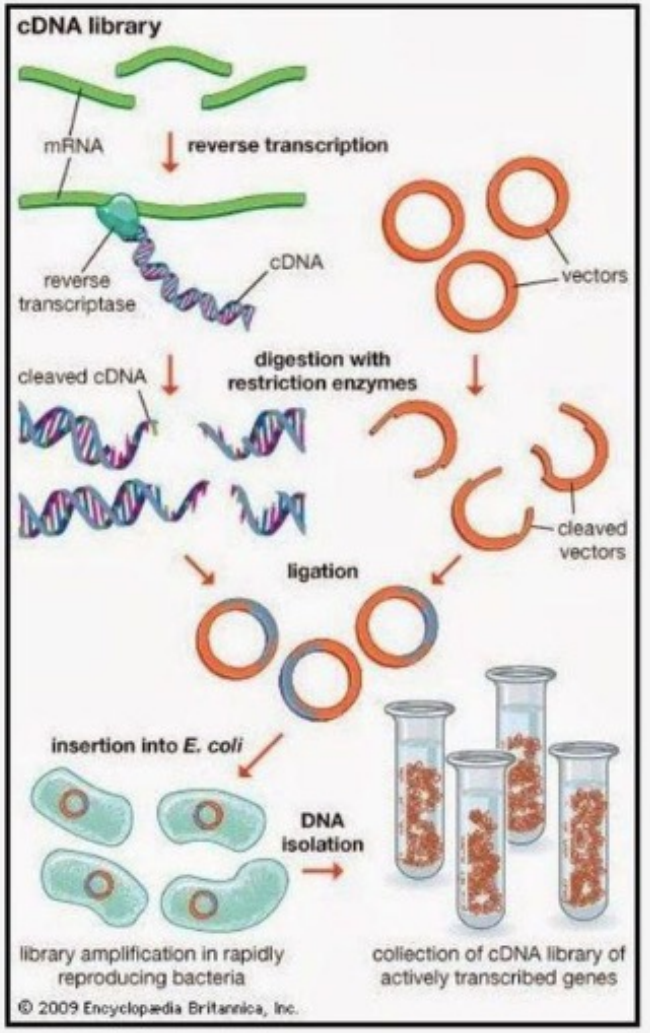
WHAT ARE ESTs?

- Expressed mRNA reverse transcribed to cDNA
- Short (200-800 bp) single-pass partial sequence reads of a mRNA pool
- Unedited and randomly selected from cDNA libraries
- Can be sequenced from either 5' or 3' end of cDNA
  - **5'-EST** generally represent **coding sequences** of mRNA (**more conserved**)
  - **3'-EST** generally represent the **3'-untranslated region** of mRNA (**less conserved**)
- Quick and cost-effective method to find insight about the diversity of genes expressed.



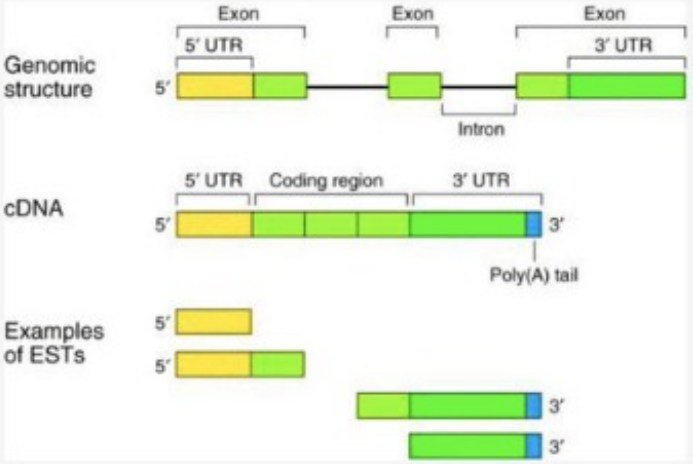
cDNA LIBRARIES

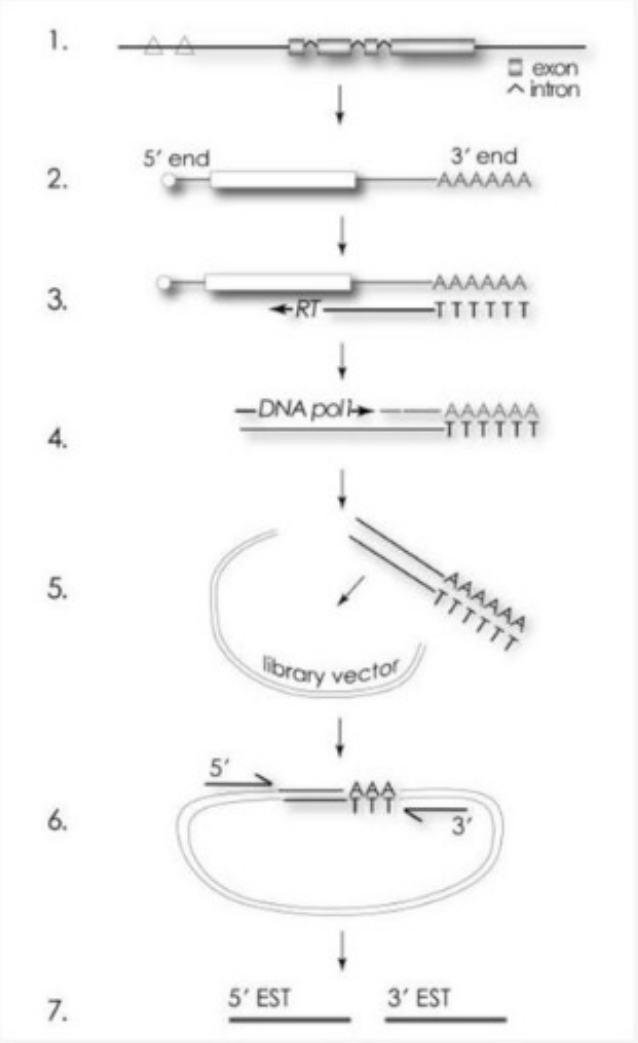
- mRNA is **unstable** and **degrades quickly** outside the cell, must be reverse transcribed into more stable cDNA
- Library represents **snapshot** of **transcribed genes** at moment of sampling
- **Populations**: single cell, tissue, organism, pooled populations
- **Stages** to consider: normal homeostasis, growth and development, pathogen defense etc…
- **Normalization** and **Subtraction** protocols available to decrease redundancy (e.g., comparable representation of all expressed mRNAs)



GENERATION OF ESTs

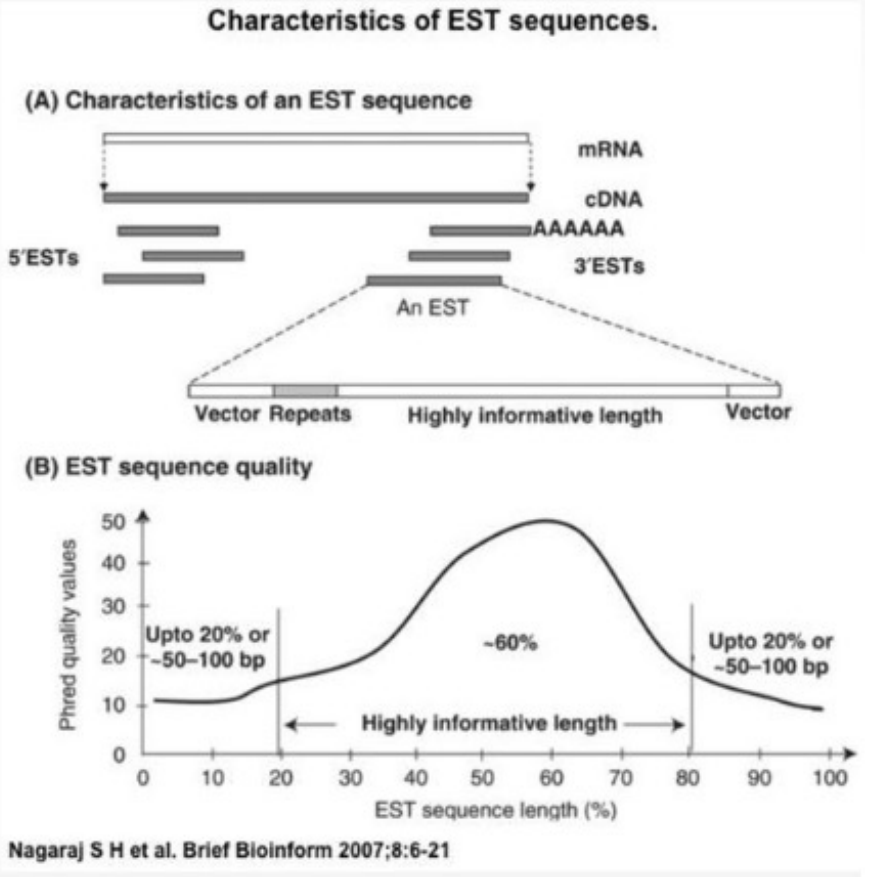
- Cloned cDNA are randomly sequenced from both 5' and 3' directions in a single pass without validation
- Length of cDNA template can vary resulting in redundant ESTs
- Methods available for producing primers that hybridize to central protein coding region (ORESTES)





CHARACTERISTICS OF ESTs

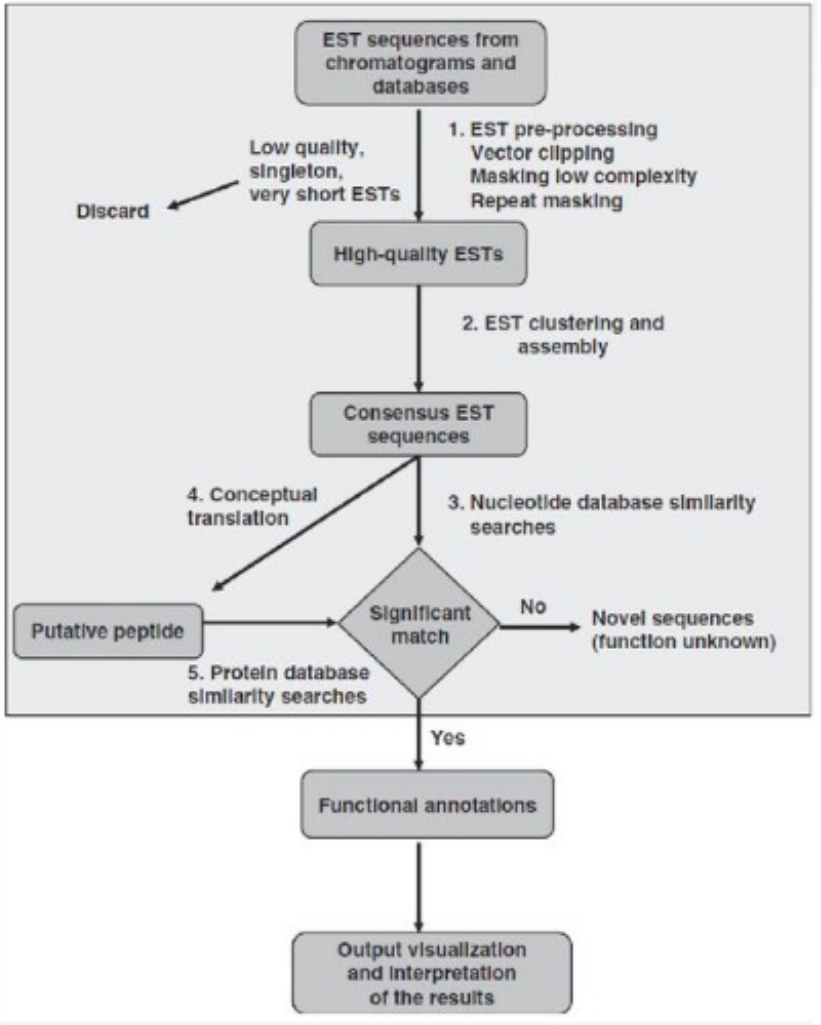
- A single raw EST has negligible biological information (low fidelity) → Power in numbers
- Single-pass sequencing results in ~3% **base-calling error rate**
- Subject to **sampling bias**: poorly expressed genes will be poorly represented
- Prevalent **vector contamination**, often transcribed along with EST sequence
- **Repetitive elements** can lead to erroneous sequence assembly
- **Poly(A)** not part of genomic DNA



EST PROCESSING PIPELINE

Generic steps in EST analysis

- **Pre-processing**: Reduces overall noise by removing vector fragments and masking repeats
- **Clustering**: Reduces redundancy by associating overlapping ESTs from a transcript of a single gene into a unique cluster (fragmented).
- **Assembling**: Generated consensus sequence (representing a putative gene) is generated by joining clusters
- **Conceptual translation**: Translation of a consensus sequence to a putative peptide



EST CLUSTERING: TWO APPROACHES

- **Stringent Clustering**: results in generally accurate clusters by generating shorter consensus sequences but with low coverage of expressed genes.

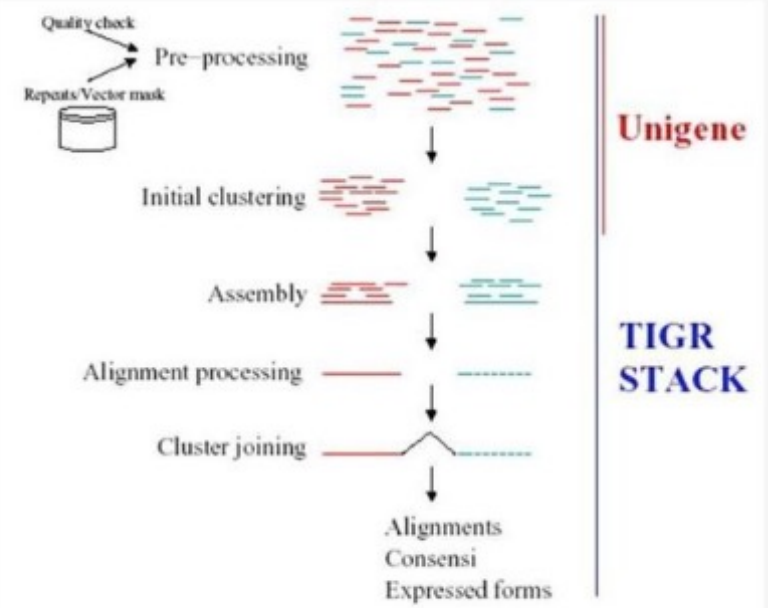
- **Loose clustering**: results in longer but less accurate consensus sequences with better coverage of gene data and alternatively spliced transcripts at the risk of including paralogues (two genes that diverge after a duplication event).

EST DATABASES

- Largest Database: **NCBI dbEST** (A descriptive catalog of ESTs)
- Number of public entries: 74,186,692 (2013)
- ESTs submitted are automatically screened and annotated
- Holds not only human ESTs but 300+ other organisms as well

Other specialized databases:

- **UniGene** database at NCBI stores unique genes and represents a non-redundant set of gene-oriented clusters generated from ESTs.
- The **TIGR** Gene Indices use **stringent clustering** approach to generate longer, virtual transcripts.
- **STACK** uses tissue-specific classification for **loose clustering** and assembly of ESTs.



APPLICATIONS: GENE DISCOVERY

- Transcriptome reconstruction: gene discovery using expression evidence.
- Not guaranteed to identify entire gene space—genes not transcribed during library construction will be missed.
- Example: sequence conservation within and between taxa

A transcriptomic analysis of the phylum Nematoda

John Parkinson[1,2], Makedonka Mitreva[3], Claire Whitton[2], Marian Thomson[2], Jennifer Daub[2], John Martin[3], Ralf Schmid[2], Neil Hall[4,6], Bart Barrell[4], Robert H Waterston[3,6], James P McCarter[3,5] & Mark L Blaxter[2]

The phylum Nematoda occupies a huge range of ecological niches, from free-living microbivores to human parasites. We analyzed the genomic biology of the phylum using 265,494 expressed-sequence tag sequences, corresponding to 93,645 putative genes, from 30 species, including 28 parasites. From 35% to 70% of each species' genes had significant similarity to proteins from the model nematode *Caenorhabditis elegans*. More than half of the putative genes were unique to the phylum, and 23% were unique to the species from which they were derived. We have not yet come close to exhausting the genomic diversity of the phylum. We identified more than 2,600 different known protein domains, some of which had differential abundances between major taxonomic groups of nematodes. We also defined 4,228 nematode-specific protein families from nematode-restricted genes: this class of genes probably underpins species- and higher-level taxonomic disparity. Nematode-specific families are particularly interesting as drug and vaccine targets.

APPLICATIONS: SEQUENCE DISCOVERY & PHYLOGENETICS

- Address questions of taxonomy and diversity.
- Explore eukaryotic (and prokaryotic) diversity at depth not possible using only (few) fully sequenced genomes.
- Example: Significant differences between prokaryotes and eukaryotes.

Research

Open Access

The global landscape of sequence diversity

José Manuel Peregrín-Álvarez*† and John Parkinson*†

Abstract

Background: Systematic comparisons between genomic sequence datasets have revealed a wide spectrum of sequence specificity from sequences that are highly conserved to those that are specific to individual species. Due to the limited number of fully sequenced eukaryotic genomes, analyses of this spectrum have largely focused on prokaryotes. Combining existing genomic datasets with the partial genomes of 193 eukaryotes derived from collections of expressed sequence tags, we performed a quantitative analysis of the sequence specificity spectrum to provide a global view of the origins and extent of sequence diversity across the three domains of life.

Results: Comparisons with prokaryotic datasets reveal a greater genetic diversity within eukaryotes that may be related to differences in modes of genetic inheritance. Mapping this diversity within a phylogenetic framework revealed that the majority of sequences are either highly conserved or specific to the species or taxon from which they derive. Between these two extremes, several evolutionary landmarks consisting of large numbers of sequences conserved within specific taxonomic groups were identified. For example, 8% of sequences derived from metazoan species are specific and conserved within the metazoan lineage. Many of these sequences likely mediate metazoan specific functions, such as cell-cell communication and differentiation.

Conclusion: Through the use of partial genome datasets, this study provides a unique perspective of sequence conservation across the three domains of life. The provision of taxon restricted sequences should prove valuable for future computational and biochemical analyses aimed at understanding evolutionary and functional relationships.

APPLICATIONS: TRANSCRIPT PROFILING

- Study of expression patterns of genes over a range of cell types, life/development stages, environmental conditions…
- Example: Identification of genes associated with tumors.

- Example: Identification of genes associated with neural development.

GEPIS—quantitative gene expression profiling in normal and cancer tissues

Yan Zhang ✉, David A. Eberhard, Gretchen D. Frantz, Patrick Dowd, Thomas D. Wu, Yan Zhou, Colin Watanabe, Shiuh-Ming Luoh, Paul Polakis, Kenneth J. Hillan, ... Show more

Bioinformatics, Volume 20, Issue 15, 12 October 2004, Pages 2390–2398, https://doi-org.libproxy.unm.edu/10.1093/bioinformatics/bch256
Published: 08 April 2004 **Article history**

Abstract

Motivation: Expression profiling in diverse tissues is fundamental to understanding gene function as well as therapeutic target identification. The vast collection of expressed sequence tags (ESTs) and the associated tissue source information provides an attractive opportunity for studying gene expression.

eXPRESSION: An in silico tool to predict patterns of gene expression

Deborah A. Ferguson[a,1], Jing-Tzyh Alan Chiang[a,1], James A. Richardson[b], Jonathan Graff[a,*]

[a]*Center for Developmental Biology, UT Southwestern Medical Center, 6000 Harry Hines Boulevard, NB5.208, Dallas, TX 75390-9133, USA*
[b]*Department of Pathology, UT Southwestern Medical Center, 6000 Harry Hines Boulevard, NB6.420, Dallas, TX 75390-9133, USA*

Received 4 October 2004; received in revised form 17 January 2005; accepted 9 March 2005
Available online 19 April 2005

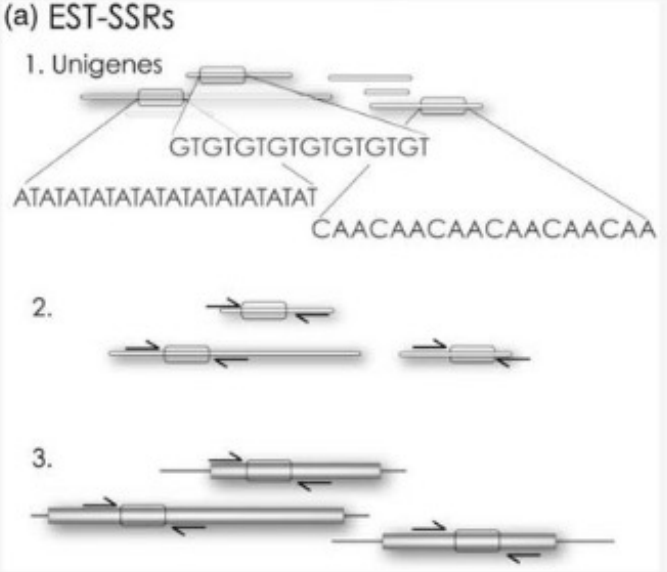
Abstract

In embryological studies, expression pattern analyses are of special importance since genes that have temporally and spatially restricted expression are not only essential as lineage markers but are often causative in formation of specific fates. Further, where a molecule is expressed can be quite revealing in regard to its endogenous function. We present a gene discovery tool, termed eXPRESSION, that utilizes the public EST databases to identify genes matching desired transcriptional profiles. We first tested and validated the ability of eXPRESSION to discover tissue-specific genes in the adult mouse; empirically as well as with DNA microarrays and RT-PCRs. These studies showed that eXPRESSION predictions could identify genes that are specifically expressed in adult mouse tissues. Next, we developed a novel search strategy to find genes that are expressed in specific regions or tissues of the developing mouse embryo. With these tools, we identified several novel genes that exhibited a neural-specific or neural-enriched expression pattern during murine development. The data show that eXPRESSION is widely applicable and may be used to identify both adult and embryonic tissue- or organ-specific genes with minimal cost and effort.

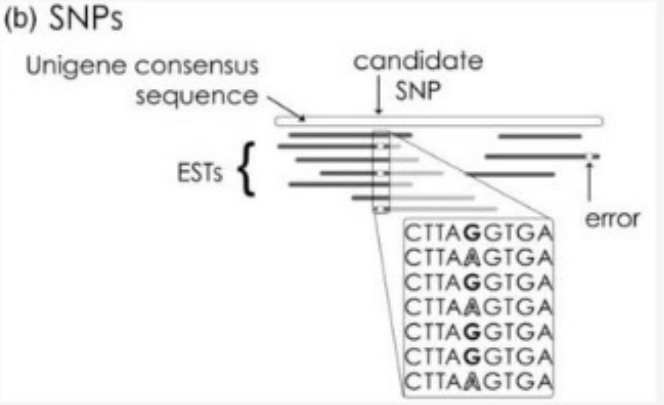
MORE APPLICATIONS OF ESTs

EST-SSRs.

1. Scan batches of unigenes for **SSR**s (or microsatellites) using software
2. Primers (horizontal arrows) are designed from unigene sequences flanking SSRs
3. which are then used for genotyping.



SNPs are identified directly from alignments of ESTs sequenced from different alleles, based on the occurrence of the same base call discrepancy in multiple sequences. Discrepancies occurring only once are likely to be sequencing errors.



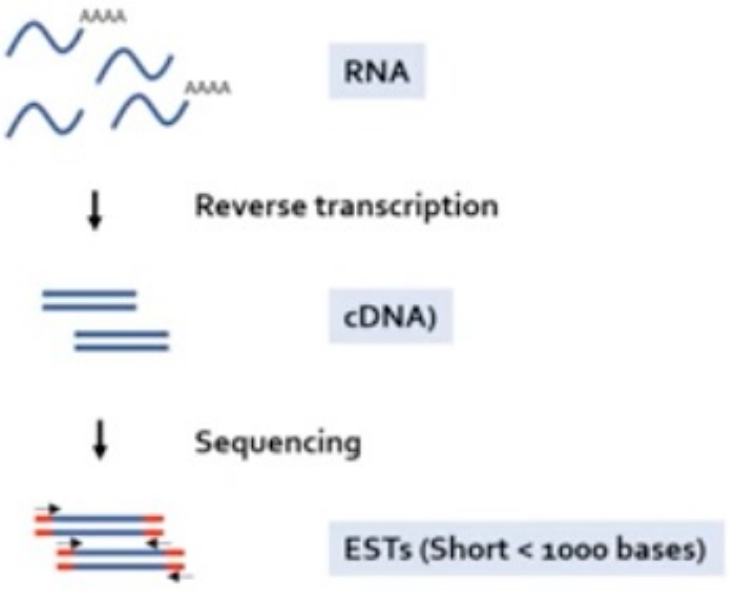
REVIEW: APPLICATIONS OF ESTs

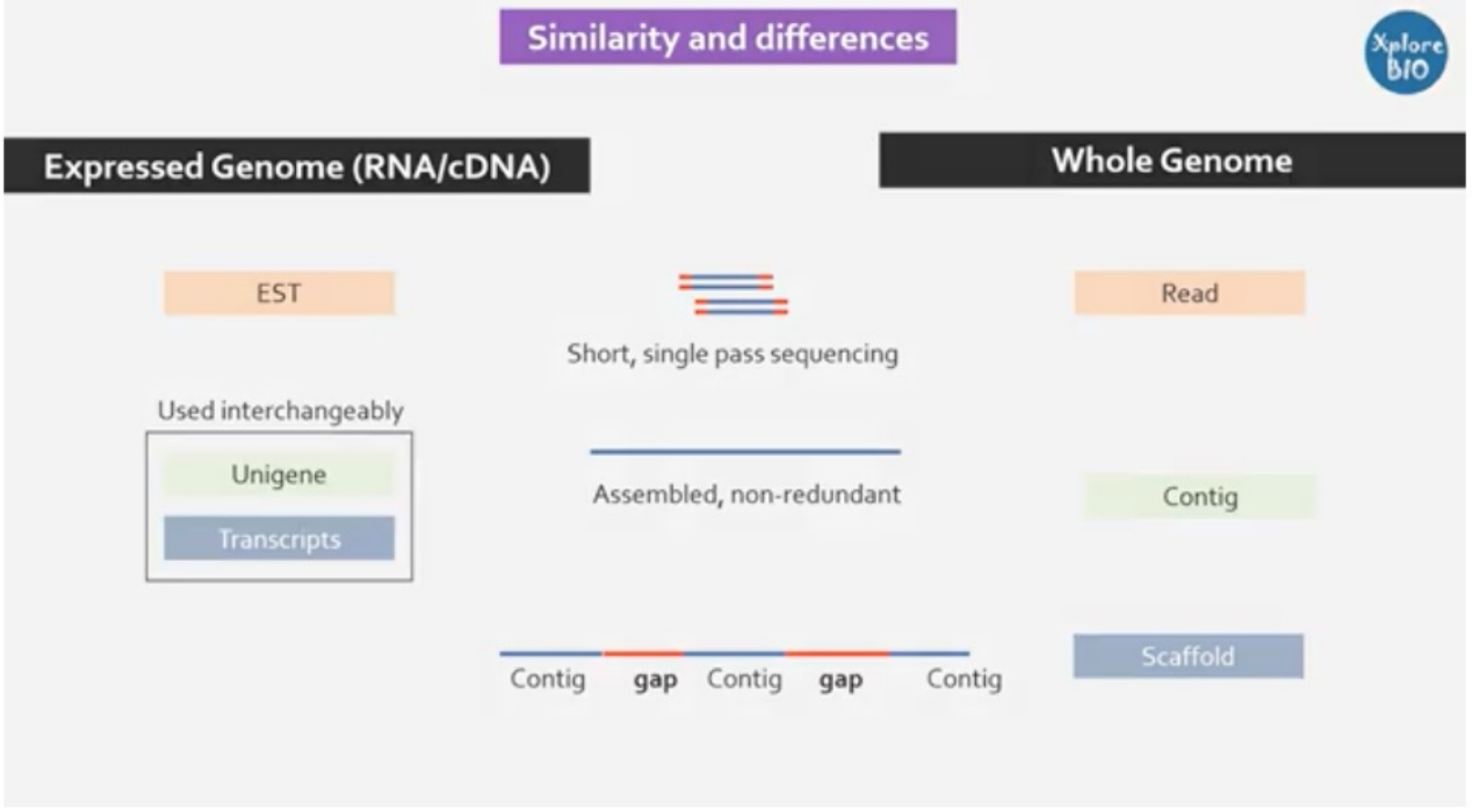
- Mapping of gene-based site markers using Sequence-Tagged Sites (STSs) derived from ESTs
- Gene structure prediction
- Investigate alternative splicing
- Discriminate between genes exhibiting tissue or disease-specific expression
- Identification and analysis of coexpressed genes on a large scale
- ESTs are also a useful resource for designing probes for DNA microarrays used to determine gene expression.
- Gene boundaries have been predicted using poly(A) sites from EST clusters

ESTs (Expressed Sequence tags)

Key features

- Expressed genome
- Stage, tissue, time specific
- Lack introns
- May be protein coding or not
- Short sequence
- No assembly
- dbEST database of GenBank
- Full length genes not obtained
- Redundancy




Similarity and differences
Xplore BIO
Expressed Genome (RNA/cDNA)
Whole Genome
EST
Read
Short, single pass sequencing
Used interchangeably
Unigene
Transcripts
Assembled, non-redundant
Contig
Scaffold
Contig
gap
Contig
gap
Contig

Why important ?

- SNPs that are not in protein coding regions may still have consequences for gene splicing, transcription factor binding, or the sequence of non-coding RNA.

- SNPs in humans can affect how humans develop diseases, respond to pathogens, chemicals, drugs, etc.

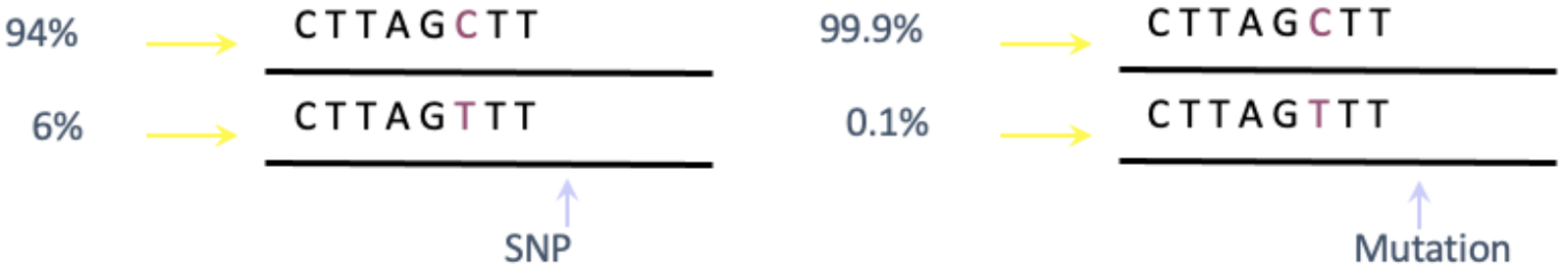
- SNPs are inherited and do not change much from generation to generation in an individual with time,
  - SNPs are of great value to biomedical research and in developing diagnostic and pharmaceutical products.

Genetic Variations

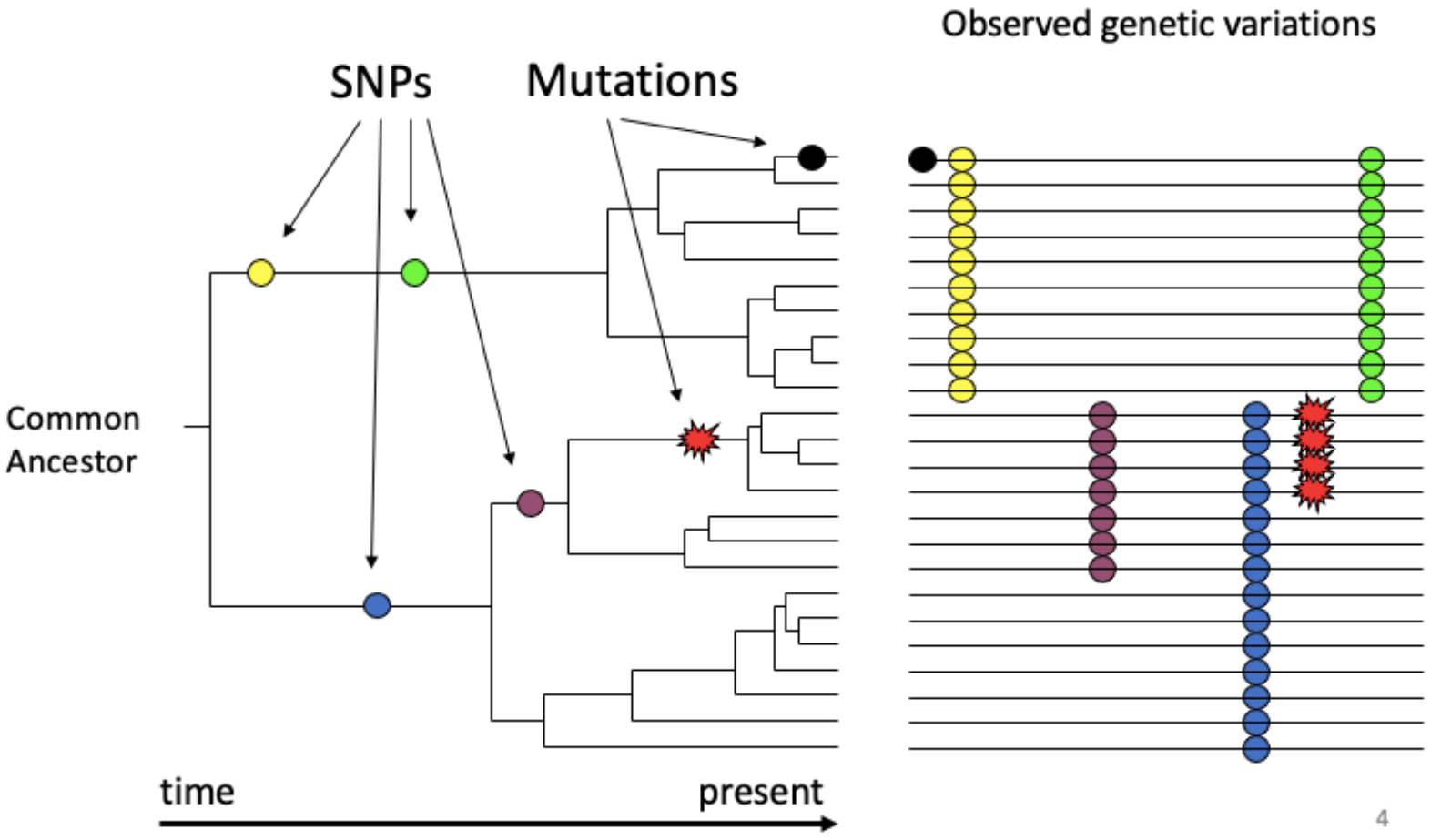
- The genetic variations in DNA sequences (e.g., insertions, deletions, and mutations) have a major impact on genetic diseases and phenotypic differences.
  - All humans share 99% the same DNA sequence.
  - The genetic variations in the coding region may change the codon of an amino acid and alters the amino acid sequence.

Single Nucleotide Polymorphism

- A Single Nucleotide Polymorphisms (SNP), pronounced "snip," is a genetic variation when a single nucleotide (i.e., A, T, C, or G) is altered and kept through heredity.
  - **SNP: Single DNA base variation found >1%**
  - **Mutation: Single DNA base variation found <1%**



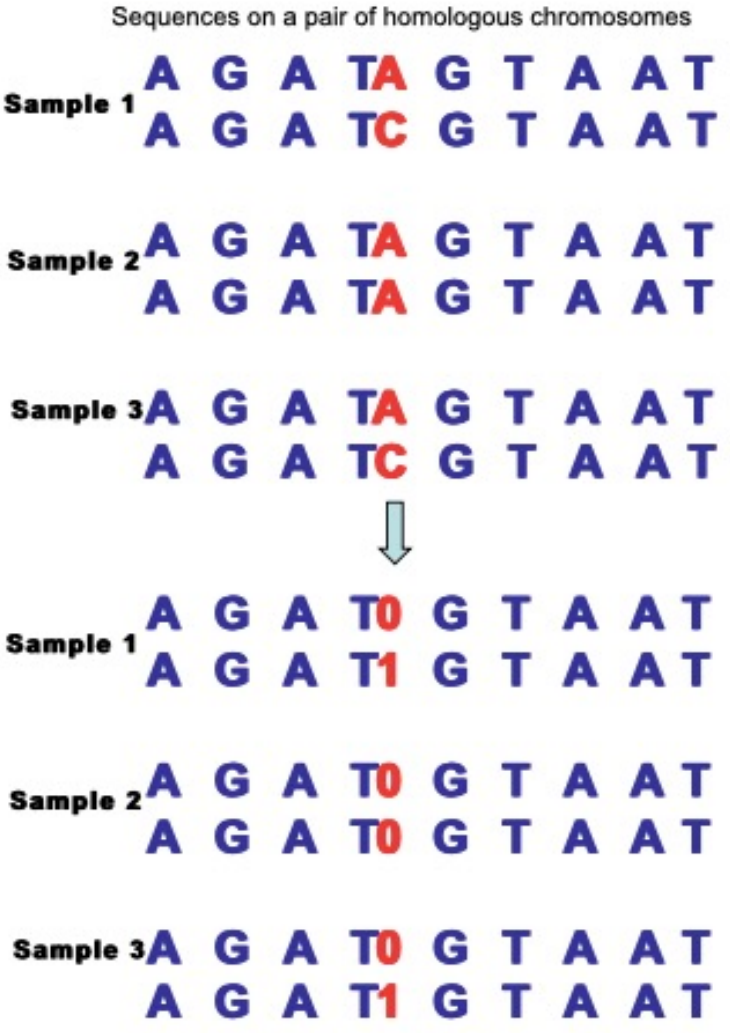
Mutations and SNPs



Single Nucleotide Polymorphism

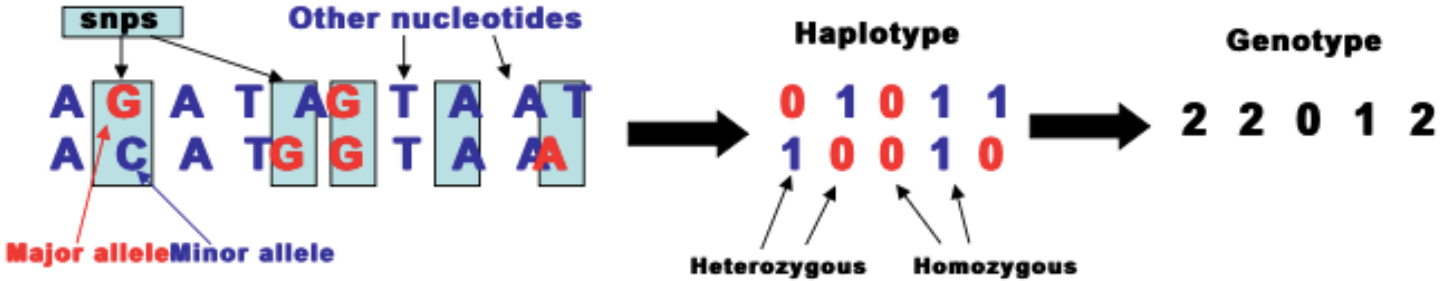
- SNPs are the most frequent form among various genetic variations.
  - 90% of human genetic variations come from SNPs.
  - SNPs occur about every 300~600 base pairs.
  - Millions of SNPs have been identified (e.g., HapMap and Perlegen).
- SNPs have become the preferred markers for association studies because of their high abundance and high-throughput SNP genotyping technologies.

- Assumption: a snp is bi-allelic.
- Major allele
  - most frequent allele
- Minor allele
  - The other one
- Example
  - Given DNA sequence
    - Major allele (A) - 67%
    - Minor allele (C) - 33%
- Encoding
  - Major allele : 0
  - Minor allele : 1



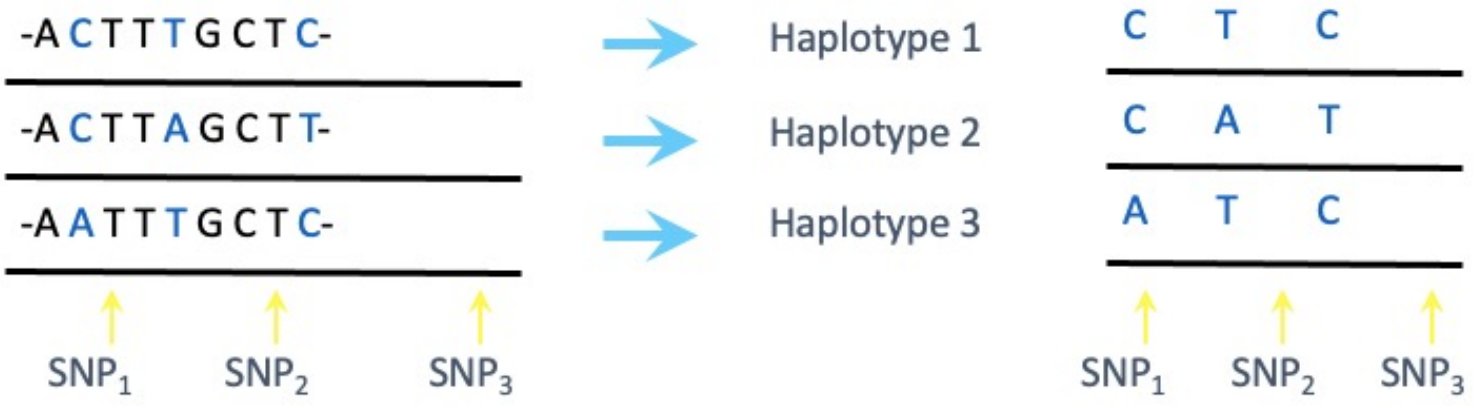
Haplotypes and Genotypes

- **Diploid organisms**: cells have two homologous set of chromosomes.
- **Haplotype**: description of SNP alleles on a single chromosome
  - 0/1 vector, e.g., 00110101 (here, 0 is for major, 1 is for minor allele).
- **Genotype**: combined description of SNP alleles on pairs of homologous chromosomes
  - 0/1/2 vector, e.g., 01122110 (0=0+0, 1=1+1, 2=0+1 or 1+0)
  - Each genotype with k 2's (heterozygotes) can be explained by 2^{k-1} pairs of haplotypes



Haplotypes

- A haplotype stands for a set of linked SNPs on the same chromosome.
  - **A haplotype can be simply considered as a binary string since each SNP is binary.**



Single Nucleotide Polymorphism

- A SNP is usually assumed to be a binary variable.
  - The probability of repeat mutation at the same SNP locus is quite small.
  - The tri-allele cases are usually considered to be the effect of genotyping errors.
- The nucleotide on a SNP locus is called
  - a major allele (if allele frequency > 50%), or
  - a minor allele (if allele frequency < 50%).

Haplotype Estimation

- Each individual has two “copies” of each chromosome.
- At each site, each chromosome has one of two alleles (states) denoted by 0 and 1 (0 major allele, 1 = minor allele)

0 1 1 1 0 0 1 1 0

1 1 0 1 0 0 1 0 0

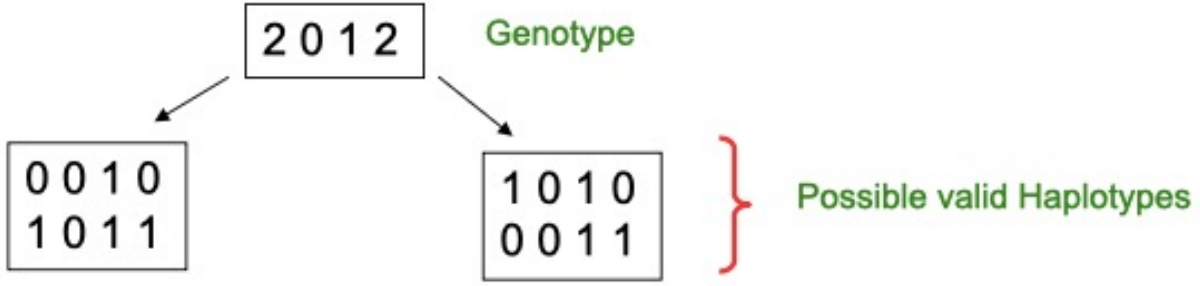
Two haplotypes per individual

Merge the haplotypes

2 1 2 1 0 0 1 2 0 Genotype for the individual

- Biological Problem: For disease association studies, haplotype data is more valuable than genotype data, but haplotype data is hard to collect. Genotype data is easy to collect.

- Computational Problem: Given a set of n genotypes, determine the original set of n haplotype pairs that generated the n genotypes.

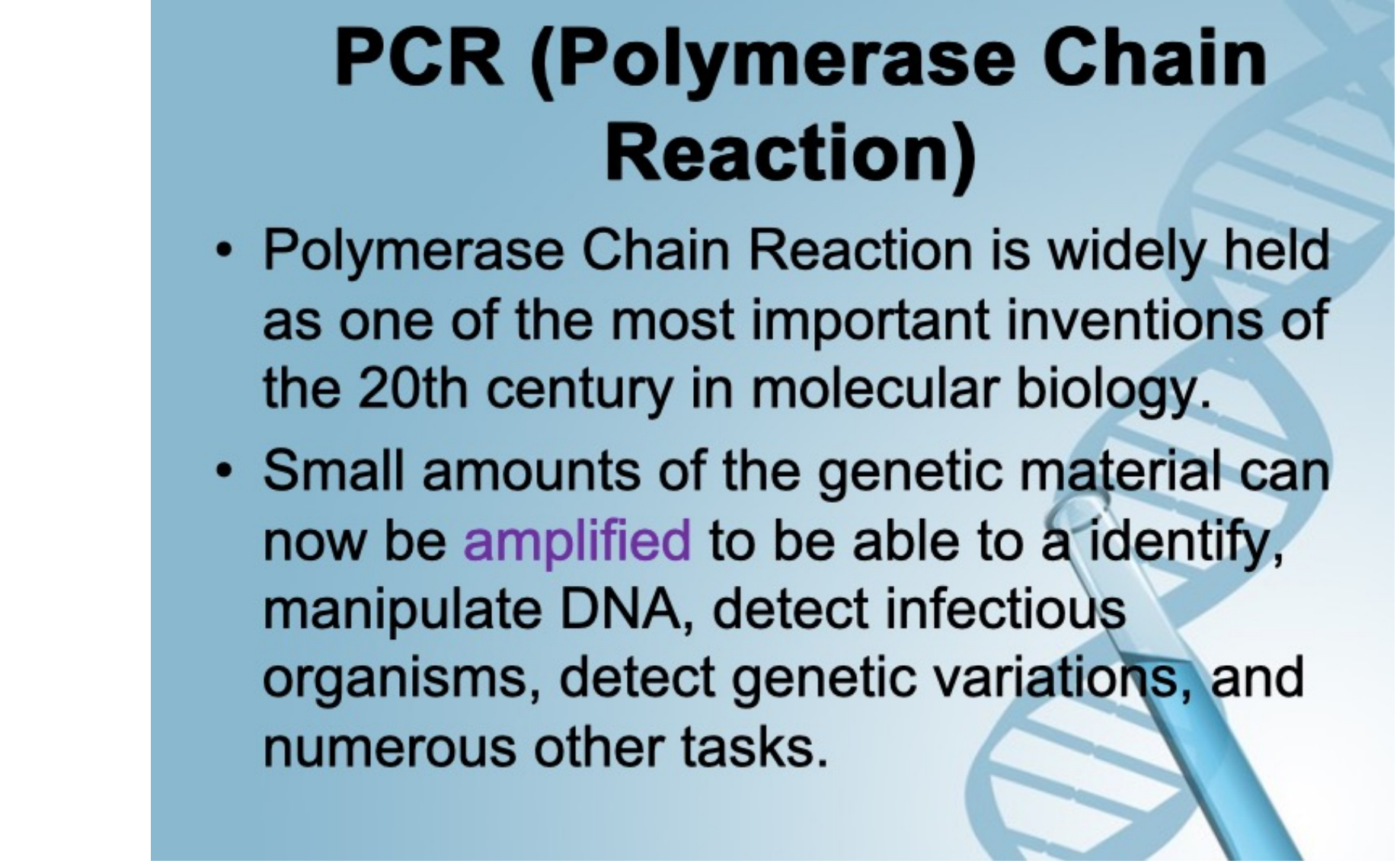


SNP databases

- HapMap project (www.hapmap.org)
  - The aim of the project is to record the significant SNPs.
  - Started in October 2002.
  - Phase 1 data have been published and analysis of Phase 2 data is underway as of October 2006.
- dbSNP
  - A database of SNPs and short deletion and insertion polymorphisms at NCBI.
- CGAP
  - Genetic variation in genes important in cancer (At the National Cancer Institute)
- EnsEMBL
  - Joint project between EMBL-EBI and the Sanger Centre to develop a system which produces and maintains automatic annotation on eukaryotic genomes.
- The SNP Consortium
  - Information about up to 300000 SNPs.

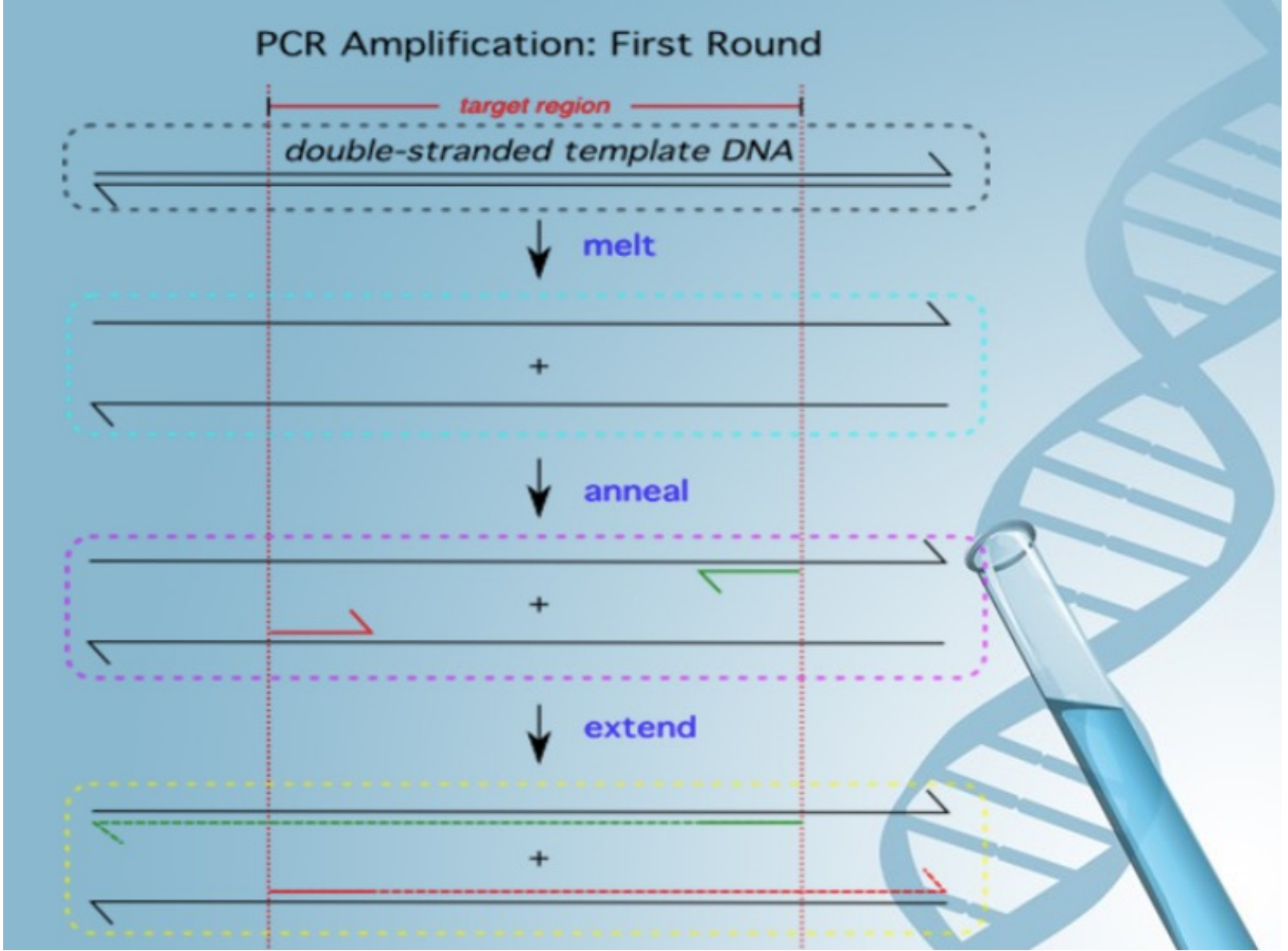
PCR (Polymerase Chain Reaction)

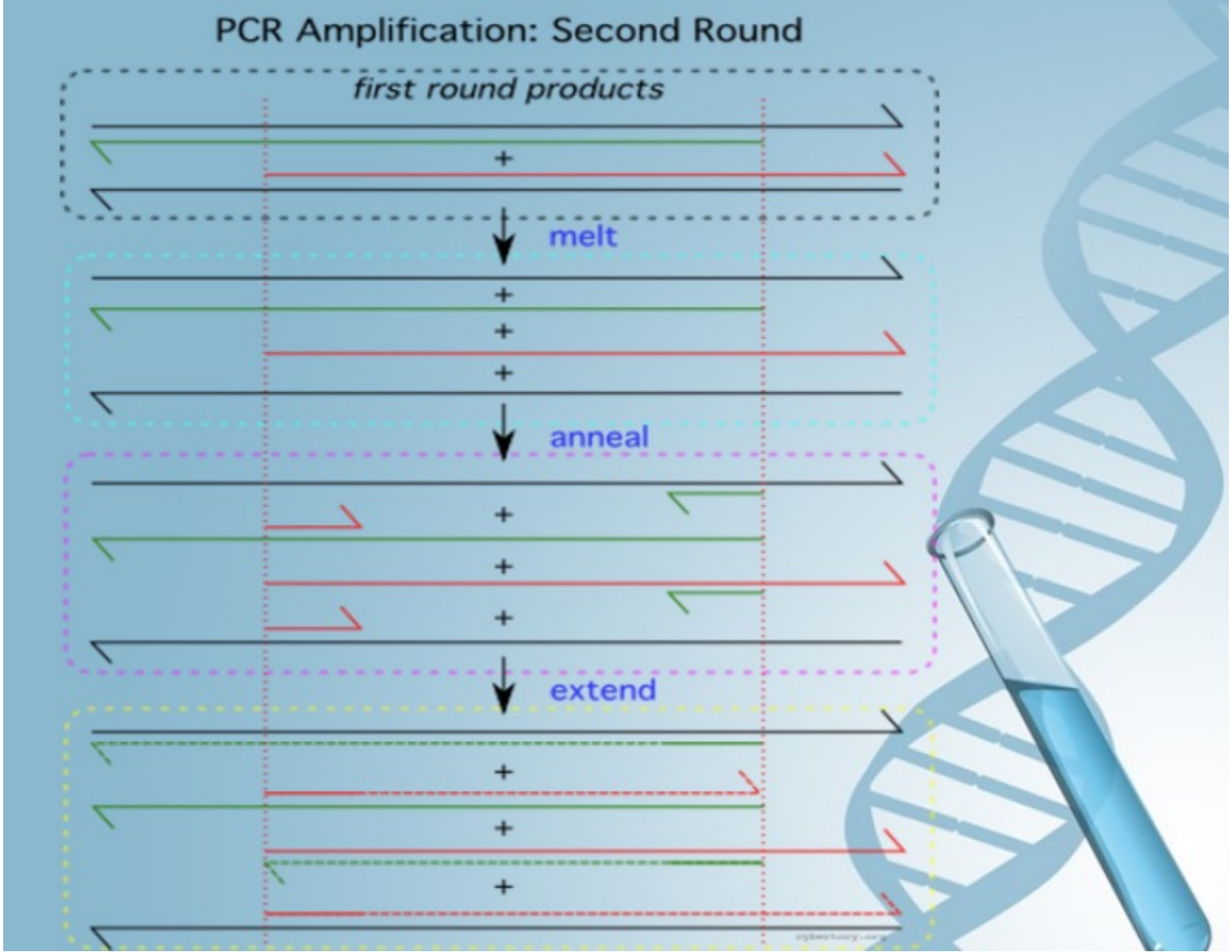
- Polymerase Chain Reaction is widely held as one of the most important inventions of the 20th century in molecular biology.
- Small amounts of the genetic material can now be amplified to be able to a identify, manipulate DNA, detect infectious organisms, detect genetic variations, and numerous other tasks.

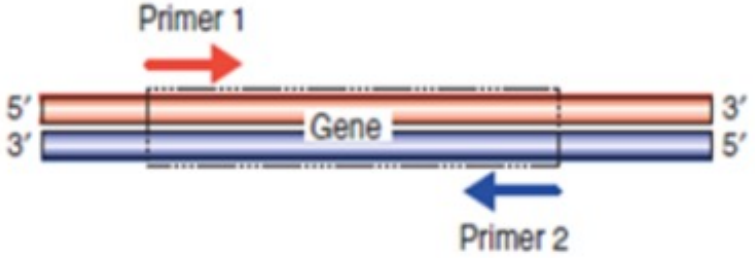


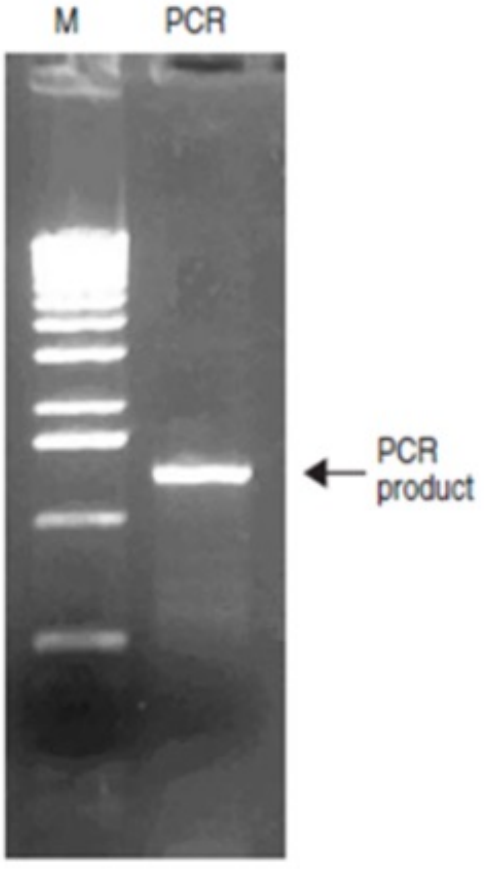
PCR involves the following three steps

- Denaturation (94°C): First, the genetic material is denatured, converting the double stranded DNA molecules to single strands.
- Annealing (60°C): The primers are then annealed to the complementary regions of the single stranded molecules.
- Extension (72°C): they are extended by the action of the DNA polymerase.


PCR Amplification: First Round
target region
double-stranded template DNA
melt
+
anneal
+
extend
+

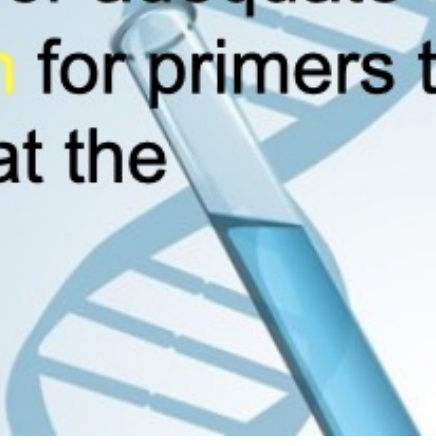

PCR Amplification: Second Round
first round products
melt
anneal
extend


Primer 1
5′
3′
Gene
3′
5′
Primer 2


M
PCR
PCR
product

Important design considerations

- **1. Primer Length:** It is generally accepted that the optimal length of PCR primers is 18-22 bp.
- This length is long enough for adequate specificity and short enough for primers to bind easily to the template at the annealing temperature.

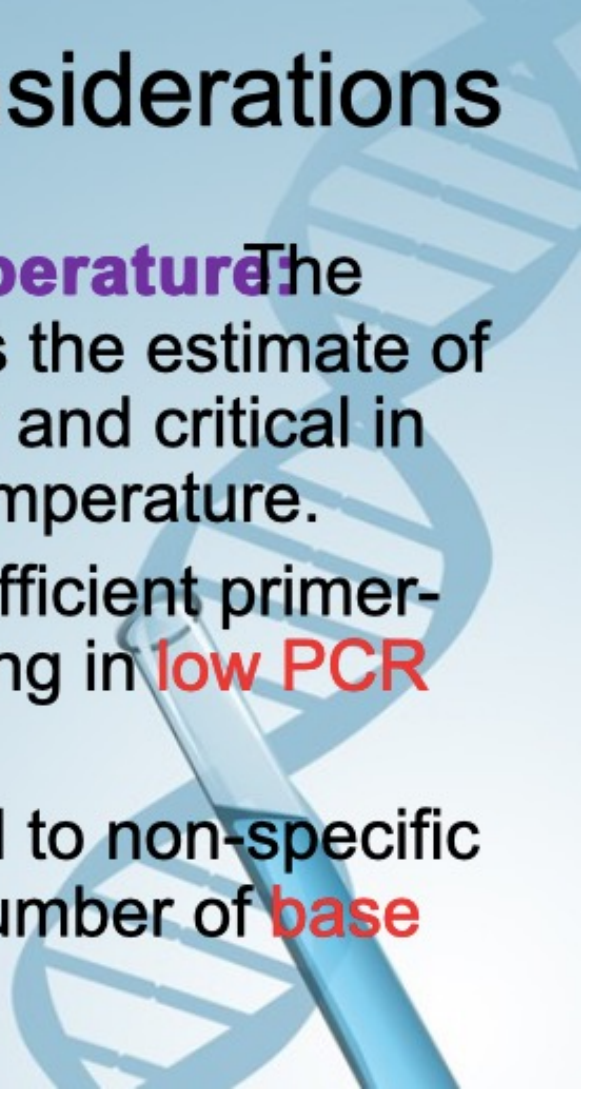


Important design considerations

- **2. Primer Melting Temperature:** This is the temperature at which 50% of the primer and its complement are hybridized.
- Primers with T_m in the range of 52-58 ^oC generally produce the best results.
- Primers with melting temperatures above 65^oC have a tendency for secondary annealing.
- The **GC content** of the sequence gives a fair indication of the primer T_m.
- "Itakura's empirical rule" makes a quick and dirty estimate of the Tm of an oligonucleotide:
  - Tm = 2 (A+T) + 4 (G+C)

Important design considerations

- **3. Primer Annealing Temperature:** The primer melting temperature is the estimate of the DNA-DNA hybrid stability and critical in determining the annealing temperature.
- Too high T_a will produce insufficient primer-template hybridization resulting in low PCR product yield.
- Too low T_a may possibly lead to non-specific products caused by a high number of base pair mismatches.



Important design considerations

- **4. GC Content:** The GC content (the number of G's and C's in the primer as a percentage of the total bases) of primer should be 40-60%.
- **5. GC Clamp:** The presence of G or C bases within the last five bases from the 3' end of primers (GC clamp) helps promote specific binding at the 3' end due to the stronger bonding of G and C bases.
  - More than 3 G's or C's should be avoided in the last 5 bases at the 3' end of the primer.

Important design considerations

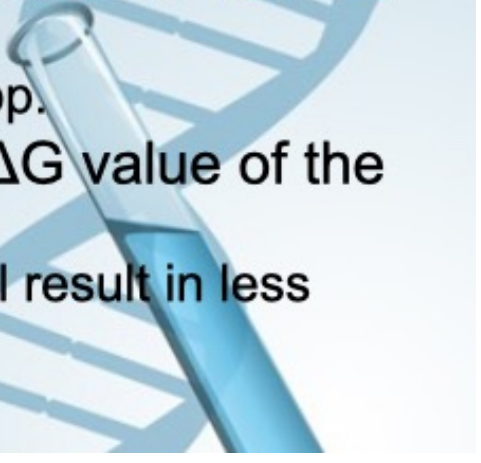
- **6. Primer Secondary Structures:**
- i) Hairpins: It is formed by intramolecular interaction within the primer.
  - Optimally a 3' end hairpin with a ΔG of -2 kcal/mol and an internal hairpin with a ΔG of -3 kcal/mol is tolerated generally.
  - ΔG definition: The Gibbs Free Energy G is the energy required to break the secondary structure.

Important design considerations

- ii) Self Dimer: A primer self-dimer is formed by intermolecular interactions between the two (same sense) primers.
  - Optimally a 3' end self dimer with a ΔG of -5 kcal/mol and an internal self dimer with a ΔG of -6 kcal/mol is tolerated generally.
- iii) Cross Dimer: Primer cross dimers are formed by intermolecular interaction between sense and antisense primers.
  - Optimally a 3' end cross dimer with a ΔG of -5 kcal/mol and an internal cross dimer with a ΔG of -6 kcal/mol is tolerated generally.

Important design considerations

- **7. Repeats:** A repeat is a di-nucleotide occurring many times consecutively. For example: ATATATAT.
  - maximum acceptable number of di-nucleotide repeats is 4.
- **8. Runs:** Primers with long runs of a single base should generally be avoided. For example, AGCGGGGGATGGGG has runs of base 'G' of value 5 and 4.
  - maximum number of runs accepted is 4 bp.
- **9. 3' End Stability:** It is the maximum ΔG value of the five bases from the 3' end.
  - An unstable 3' end (less negative ΔG) will result in less false priming.



Important design considerations

- **10. Avoid Template Secondary Structure**- The stability of the template secondary structures depends largely on their free energy and melting temperatures(T_m).
- Consideration of template secondary structures is important in designing primers, especially in qPCR.
- If primers are designed on a secondary structures which is stable even above the annealing temperatures, the primers are unable to bind to the template.
- Hence, it is important to design primers in the regions of the templates that do not form stable secondary structures.

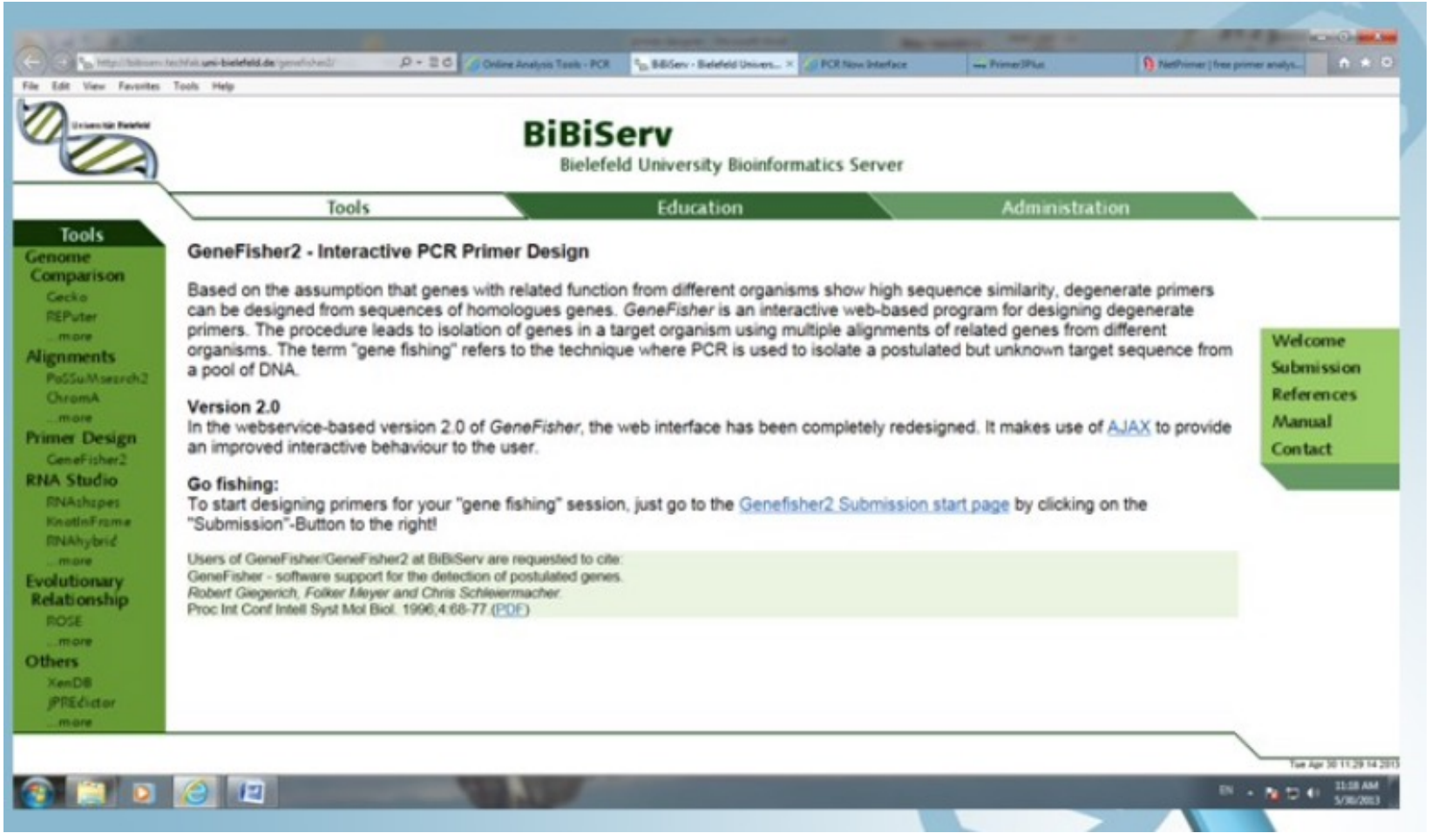
Important design considerations

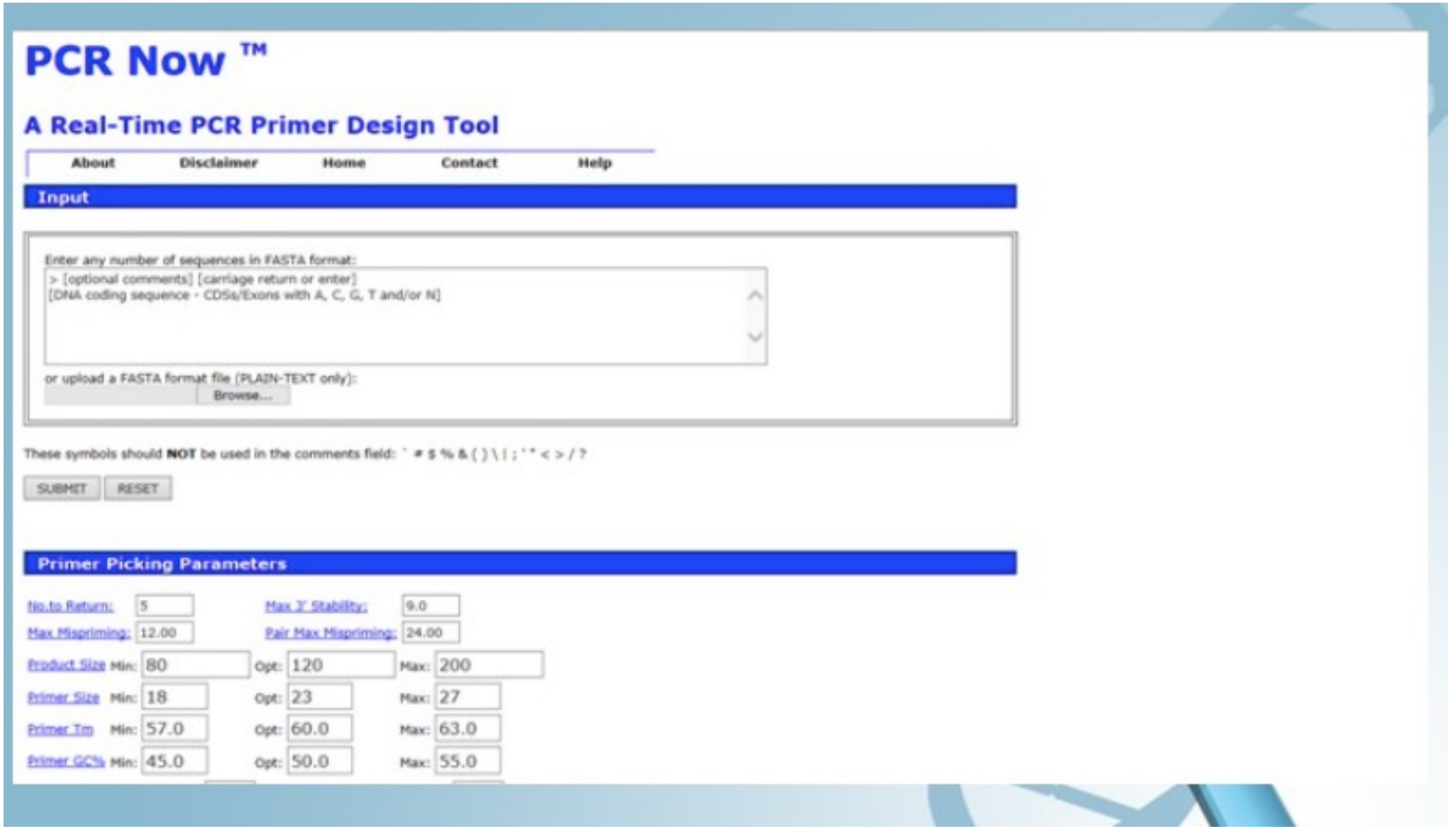
- **11. Avoid Cross Homology:** Primers designed for a sequence must not amplify other genes in the mixture.
- Commonly, primers are designed and then BLASTed to test the specificity.

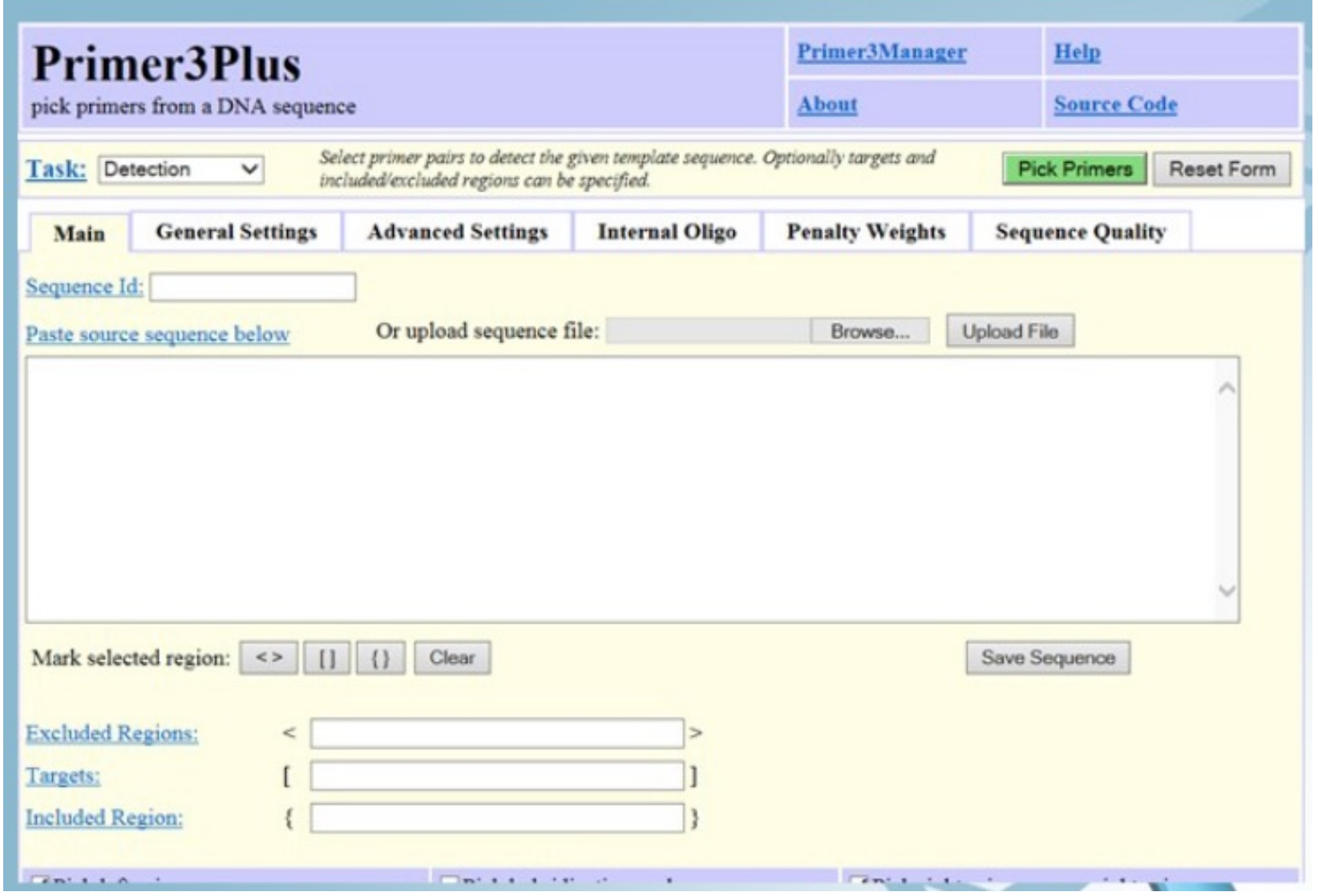
Primer Design using Software

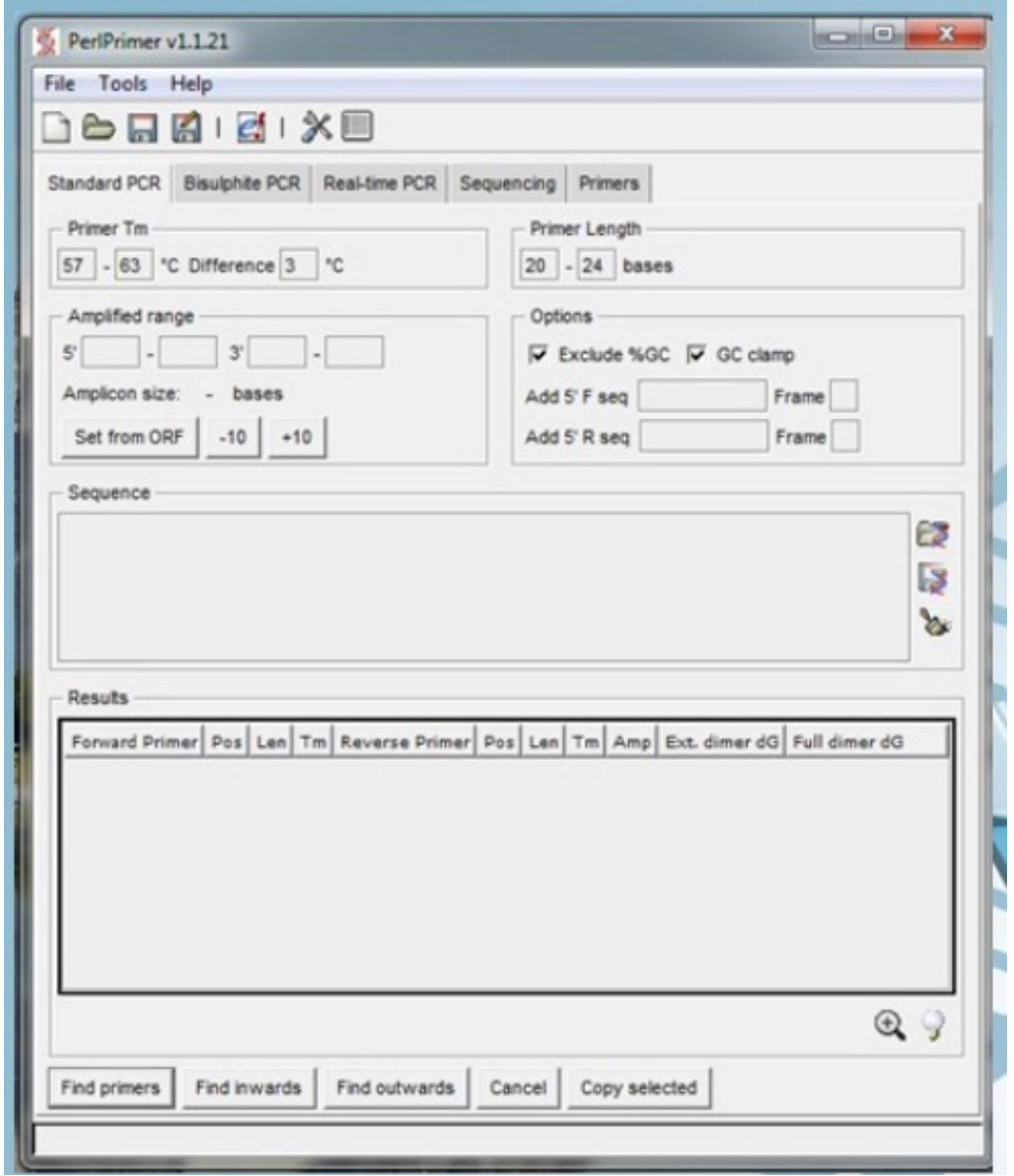
- A number of primer design tools are available that can assist in PCR primer design for new and experienced users alike.
- These tools may reduce the cost and time involved in experimentation by lowering the chances of failed experimentation.

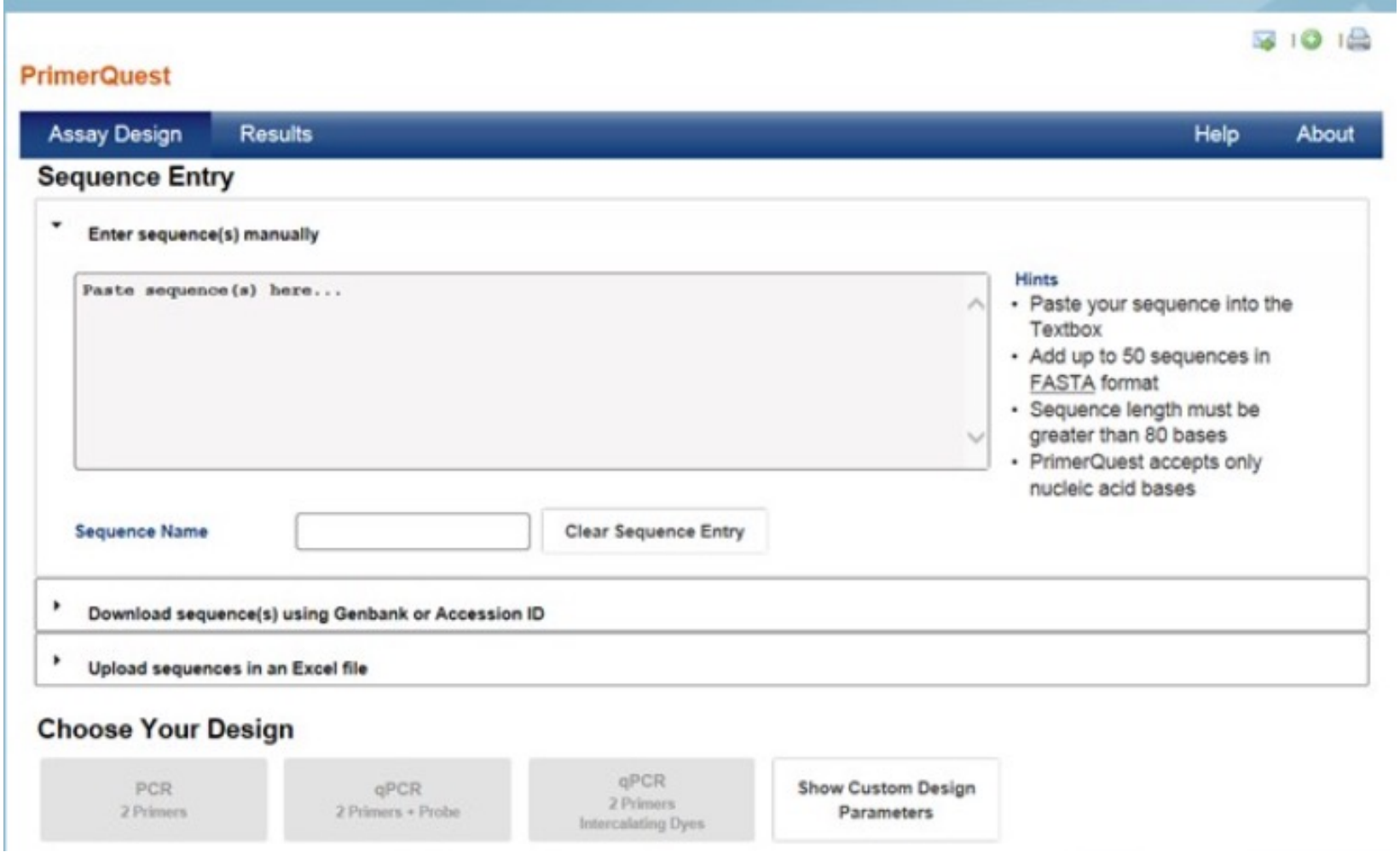
Designing Primer Softwares


File Edit View Favorites Tools Help
BiBiServ
Bielefeld University Bioinformatics Server
Tools
Education
Administration
Tools
Genome Comparison
Gecko
REPuter
...more
Alignments
PoSSuMsearch2
ChromA
...more
Primer Design
GeneFisher2
RNA Studio
RNAshapes
KnotInFrame
RNAhybrid
...more
Evolutionary Relationship
ROSE
...more
Others
XenDB
jPREdictor
...more
GeneFisher2 - Interactive PCR Primer Design
Based on the assumption that genes with related function from different organisms show high sequence similarity, degenerate primers can be designed from sequences of homologues genes. GeneFisher is an interactive web-based program for designing degenerate primers. The procedure leads to isolation of genes in a target organism using multiple alignments of related genes from different organisms. The term "gene fishing" refers to the technique where PCR is used to isolate a postulated but unknown target sequence from a pool of DNA.
Version 2.0
In the webservice-based version 2.0 of GeneFisher, the web interface has been completely redesigned. It makes use of AJAX to provide an improved interactive behaviour to the user.
Go fishing:
To start designing primers for your "gene fishing" session, just go to the Genefisher2 Submission start page by clicking on the "Submission"-Button to the right!
Users of GeneFisher/GeneFisher2 at BiBiServ are requested to cite:
GeneFisher - software support for the detection of postulated genes.
Robert Giegerich, Folker Meyer and Chris Schleiermacher.
Proc Int Conf Intell Syst Mol Biol. 1996;4:68-77 (PDF)
Welcome
Submission
References
Manual
Contact


PCR Now ™
A Real-Time PCR Primer Design Tool
About
Disclaimer
Home
Contact
Help
Input
Enter any number of sequences in FASTA format:
> [optional comments] [carriage return or enter]
[DNA coding sequence - CDSs/Exons with A, C, G, T and/or N]
or upload a FASTA format file (PLAIN-TEXT only):
Browse...
These symbols should NOT be used in the comments field: ` # $ % & { } \ | ; ' " < > / ?
SUBMIT
RESET
Primer Picking Parameters
No.to Return: 5
Max 3' Stability: 9.0
Max Mispriming: 12.00
Pair Max Mispriming: 24.00
Product Size Min: 80 Opt: 120 Max: 200
Primer Size Min: 18 Opt: 23 Max: 27
Primer Tm Min: 57.0 Opt: 60.0 Max: 63.0
Primer GC% Min: 45.0 Opt: 50.0 Max: 55.0

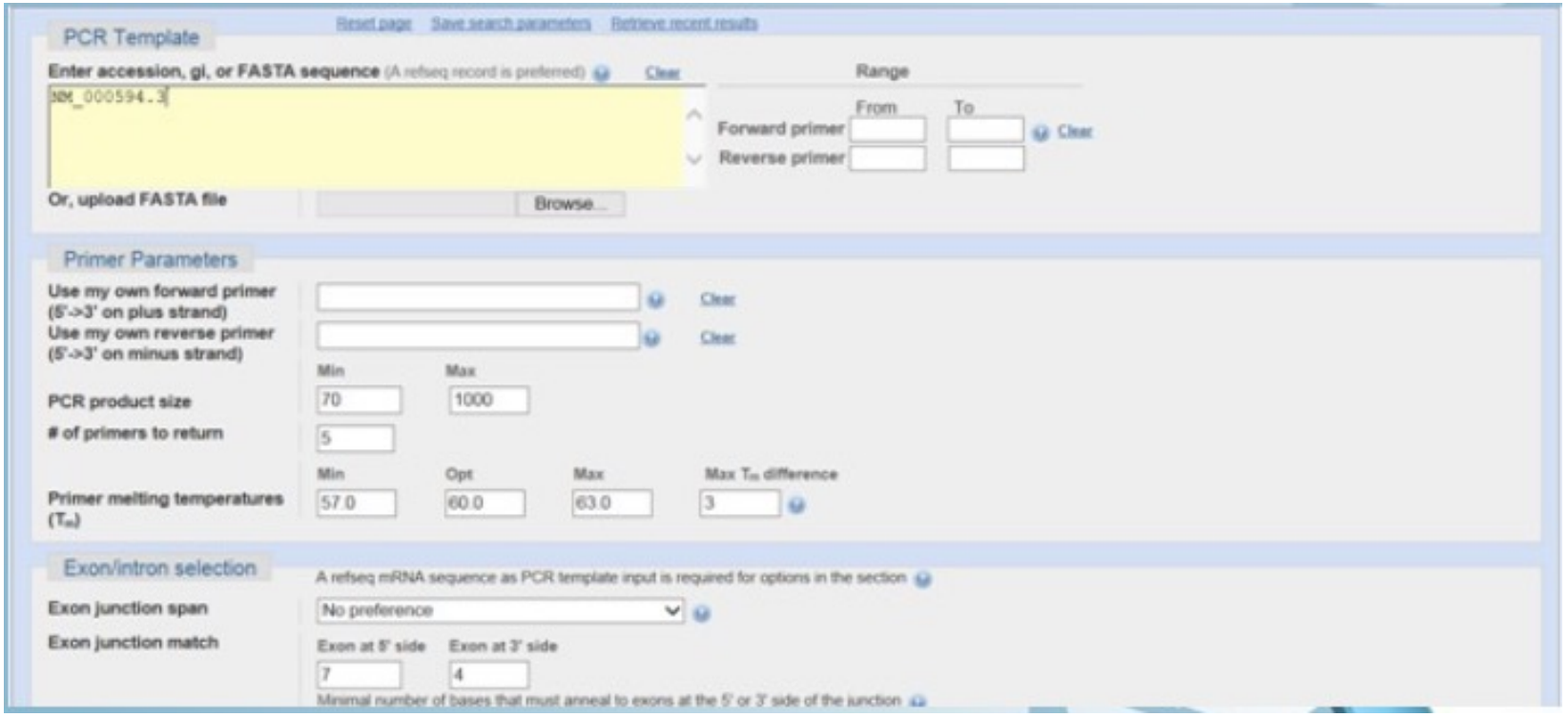

Primer3Plus
pick primers from a DNA sequence
Primer3Manager
Help
About
Source Code
Task:
Detection
Select primer pairs to detect the given template sequence. Optionally targets and included/excluded regions can be specified.
Pick Primers
Reset Form
Main
General Settings
Advanced Settings
Internal Oligo
Penalty Weights
Sequence Quality
Sequence Id:
Paste source sequence below
Or upload sequence file:
Browse...
Upload File
Mark selected region:
< >
[ ]
{ }
Clear
Save Sequence
Excluded Regions:
Targets:
Included Region:


PerlPrimer v1.1.21
File
Tools
Help
Standard PCR
Bisulphite PCR
Real-time PCR
Sequencing
Primers
Primer Tm
57 - 63 °C Difference 3 °C
Primer Length
20 - 24 bases
Amplified range
5' - 3' -
Amplicon size: - bases
Set from ORF
-10
+10
Options
Exclude %GC
GC clamp
Add 5' F seq
Frame
Add 5' R seq
Frame
Sequence
Results
Forward Primer
Pos
Len
Tm
Reverse Primer
Pos
Len
Tm
Amp
Ext. dimer dG
Full dimer dG
Find primers
Find inwards
Find outwards
Cancel
Copy selected


PrimerQuest
Assay Design
Results
Help
About
Sequence Entry
Enter sequence(s) manually
Paste sequence(s) here...
Hints
• Paste your sequence into the Textbox
• Add up to 50 sequences in FASTA format
• Sequence length must be greater than 80 bases
• PrimerQuest accepts only nucleic acid bases
Sequence Name
Clear Sequence Entry
Download sequence(s) using Genbank or Accession ID
Upload sequences in an Excel file
Choose Your Design
PCR
2 Primers
qPCR
2 Primers + Probe
qPCR
2 Primers
Intercalating Dyes
Show Custom Design Parameters

Designing Primer using NCBI


NCBI Resources How To
mghajkemi My NCBI Sign Out
Nucleotide
Nucleotide
Search
Limits Advanced
Help
Display Settings: GenBank
Send:
Change region shown
Customize view
Homo sapiens tumor necrosis factor (TNF), mRNA
NCBI Reference Sequence: NM_000594.3
FASTA Graphics
Go to:
LOCUS NM_000594 1686 bp mRNA linear PRI 19-MAY-2013
DEFINITION Homo sapiens tumor necrosis factor (TNF), mRNA.
ACCESSION NM_000594
VERSION NM_000594.3 GI:395132451
KEYWORDS RefSeq.
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 1686)
AUTHORS Sennikov,S.V., Golikova,E.A., Kireev,F.D. and Lopatnikova,J.A.
TITLE Purification of human immunoglobulin G autoantibodies to tumor
necrosis factor using affinity chromatography and magnetic
separation
JOURNAL J. Immunol. Methods 390 (1-2), 92-98 (2013)
PUBMED 23388693
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence
Articles about the TNF gene
SIRT6 regulates TNF-α secretion through
hydrolysis of long-chain fatty acyl h [Nature. 2013]
Spreading depression triggers headache by
activating neuronal Panx1 channe [Science. 2013]
Association of tumor necrosis factor-α gene G-
308A polymorphism with dis [DNA Cell Biol. 2013]
See all...


Reset page
Save search parameters
Retrieve recent results
PCR Template
Enter accession, gi, or FASTA sequence (A refseq record is preferred)
Clear
NM_000594.3
Range
From
To
Forward primer
Reverse primer
Clear
Or, upload FASTA file
Browse...
Primer Parameters
Use my own forward primer (5'->3' on plus strand)
Use my own reverse primer (5'->3' on minus strand)
Clear
Clear
Min
Max
PCR product size
70
1000
# of primers to return
5
Min
Opt
Max
Max Tm difference
Primer melting temperatures (Tm)
57.0
60.0
63.0
3
Exon/intron selection
A refseq mRNA sequence as PCR template input is required for options in the section
Exon junction span
No preference
Exon junction match
Exon at 5' side
Exon at 3' side
7
4
Minimal number of bases that must anneal to exons at the 5' or 3' side of the junction

Primer-BLAST

Primer-Blast results

NCBI Primer-BLAST : results: Job id=JSID_01_623022_130.14.22.10_9003 more...

Input PCR template	NM_000594.3 Homo sapiens tumor necrosis factor (TNF), mRNA
Range	1 - 1686
Specificity of primers	Primer pairs are specific to input template as no other targets were found in selected database: Refseq mRNA (Organism limited to Homo sapiens)
Other reports	▶Search Summary

▼ Graphical view of primer pairs

▼ Detailed primer reports

▼ Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CGAACCCCGAGTGACAAGCC	Plus	20	419	438	63.38	65.00	3.00	1.00
Reverse primer	CCATTGGCCAGGAGGGCATT	Minus	20	522	503	62.88	60.00	6.00	2.00
Product length	104								

Products on intended target

>NM_000594.3 Homo sapiens tumor necrosis factor (TNF), mRNA

```
product length = 104
Forward primer  1    CGAACCCCGAGTGACAAGCC  20
Template        419  ....................  438

Reverse primer  1    CCATTGGCCAGGAGGGCATT  20
Template        522  ....................  503
```

Primer pair 2

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AGACGCCACATCCCCTGACA	Plus	20	83	102	63.38	60.00	3.00	3.00
Reverse primer	AGCTCCACGTCCCGGATCAT	Minus	20	210	191	63.22	60.00	5.00	3.00
Product length	128								

Products on intended target

>NM_000594.3 Homo sapiens tumor necrosis factor (TNF), mRNA

```
product length = 128
Forward primer  1   AGACGCCACATCCCCTGACA  20
Template        83  ....................  102
```

Review of Proteins

- Proteins: polypeptides with a three dimensional structure
-
- **Primary structure** – sequence of amino acids constituting polypeptide chain

- **Secondary structure** – local organization of polypeptide chain into secondary structures such as α helices and β sheets

Review of Proteins

- **Tertiary structure** –three dimensional arrangements of amino acids as they react to one another due to polarity and interactions between side chains

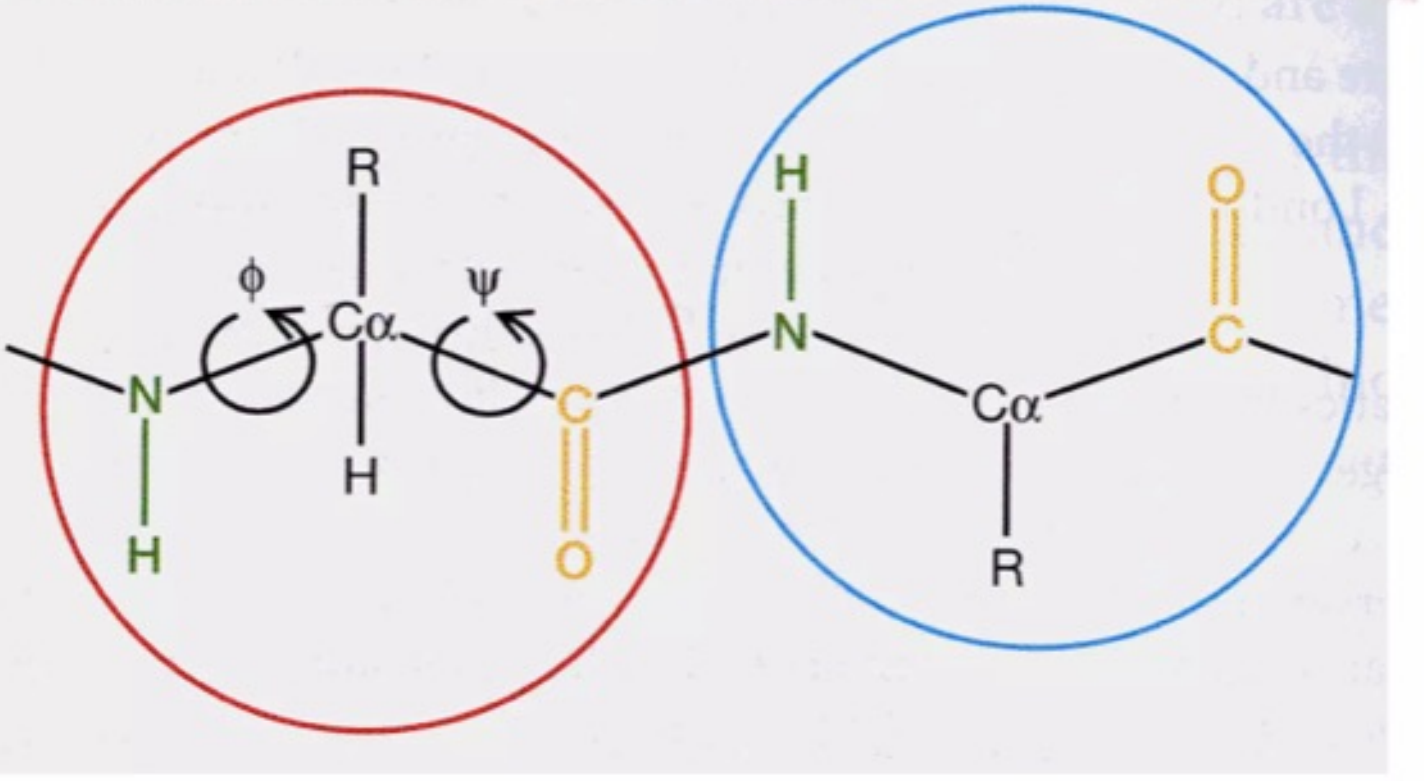
- **Quaternary structure** – Interaction of several protein subunits

Protein Structure

- Proteins: chains of amino acids joined by peptide bonds

- Amino Acids:
  - Polar (separate positive and negatively charged regions)
  - free C=O group (CARBOXYL), can act as hydrogen bond acceptor
  - free NH group (AMINYL), can act as hydrogen bond donor

Protein Structure



Protein Structure

- Many confirmations possible due to the rotation around the Alpha-Carbon (C_α) atom

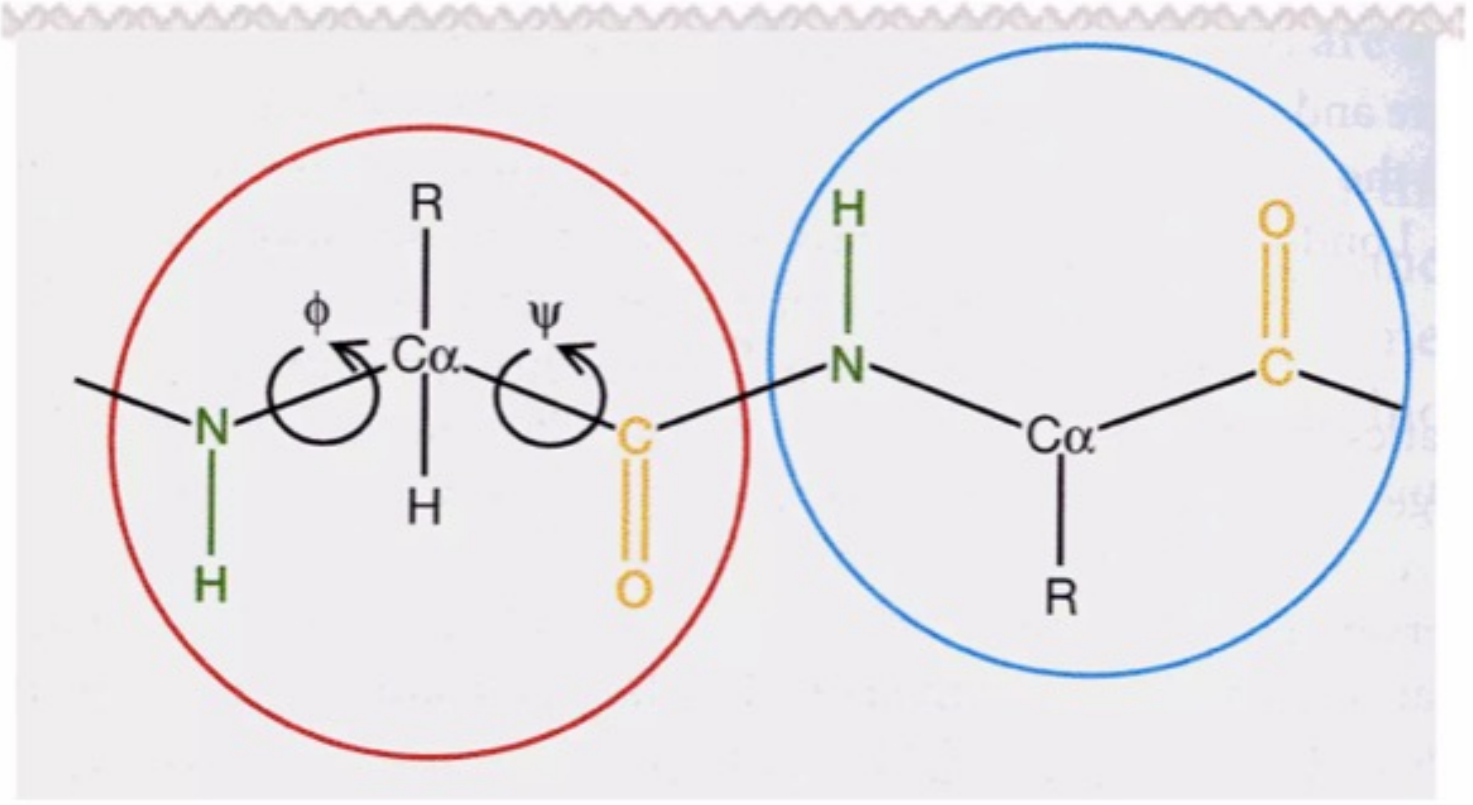
- Confirmational changes lead to differences in three-dimensional structure of protein

Protein Structure

- Polypeptide chain has pattern of N-C_α-C repeated

- Angle between aminyl group and C_α is PHI (ϕ) angle; angle between C_α and carboxyl group is PSI (ψ) angle

Protein Structure



Differences between A.A.'s

- Difference between 20 amino acids is the R side chains

- Amino acids can be separated based on the chemical properties of the side chains:
  - Hydrophobic
  - Charged
  - Polar

Differences between A.A.'s

- Hydrophobic: Alanine(A), Valine(V), phenylalanine (Y), Proline (P), Methionine (M), isoleucine (I), and Leucine(L)
- Charged: Aspartic acid (D), Glutamic Acid (E), Lysine (K), Arginine (R)
- Polar: Serine (S), Theronine (T), Tyrosine (Y); Histidine (H), Cysteine (C), Asparagine (N), Glutamine (Q), Tryptophan (W)

Secondary Structure

Secondary Structures

- Core of each protein made up of regular secondary structures
- Regular patterns of hydrogen bonds are formed between neighboring amino acids
- Amino acids in secondary structures have similar ϕ and ψ angles

Secondary Structures

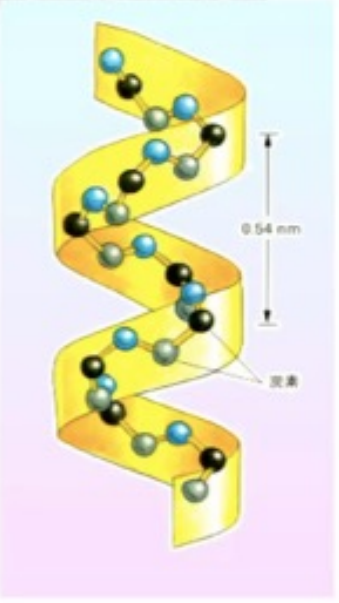
- Structures act to neutralize the polar groups on each amino acid
- Secondary structures tightly packed in protein core and a hydrophobic environment
- Each amino acid side group has a limited space to occupy -- therefore a limited number of possible interactions

Types of Secondary Structures

- α Helices
- β Sheets
- Loops
- Coils

α Helix

- Most abundant secondary structure
- 3.6 amino acids per turn
- Hydrogen bond formed between every fourth reside
- Average length: 10 amino acids, or 3 turns
- Varies from 5 to 40 amino acids



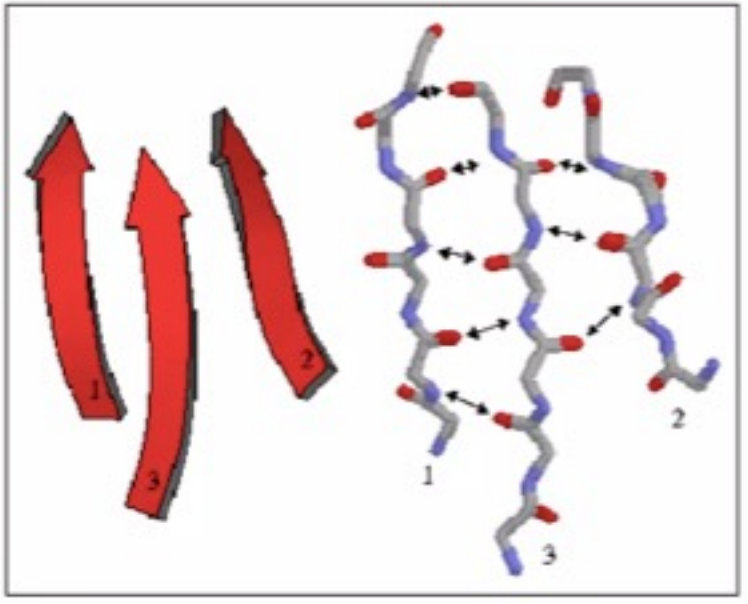
α Helix

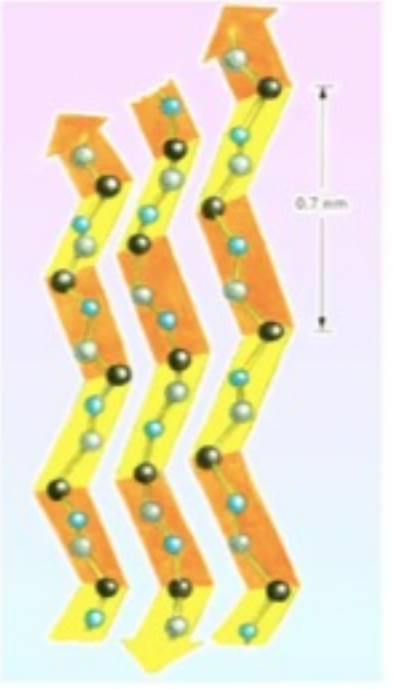
- Normally found on the surface of protein cores
- Interact with aqueous environment
  - Inner facing side has hydrophobic amino acids
  - Outer-facing side has hydrophilic amino acids

α Helix

- Every third amino acid tends to be hydrophobic
- Pattern can be detected computationally
- Rich in alanine (A), gutamic acid (E), leucine (L), and methionine (M)
- Poor in proline (P), glycine (G), tyrosine (Y), and serine (S)

β Sheet





β Sheet

- Hydrogen bonds between 5-10 consecutive amino acids in one portion of the chain with another 5-10 farther down the chain

- Interacting regions may be adjacent with a short loop, or far apart with other structures in between

β Sheet

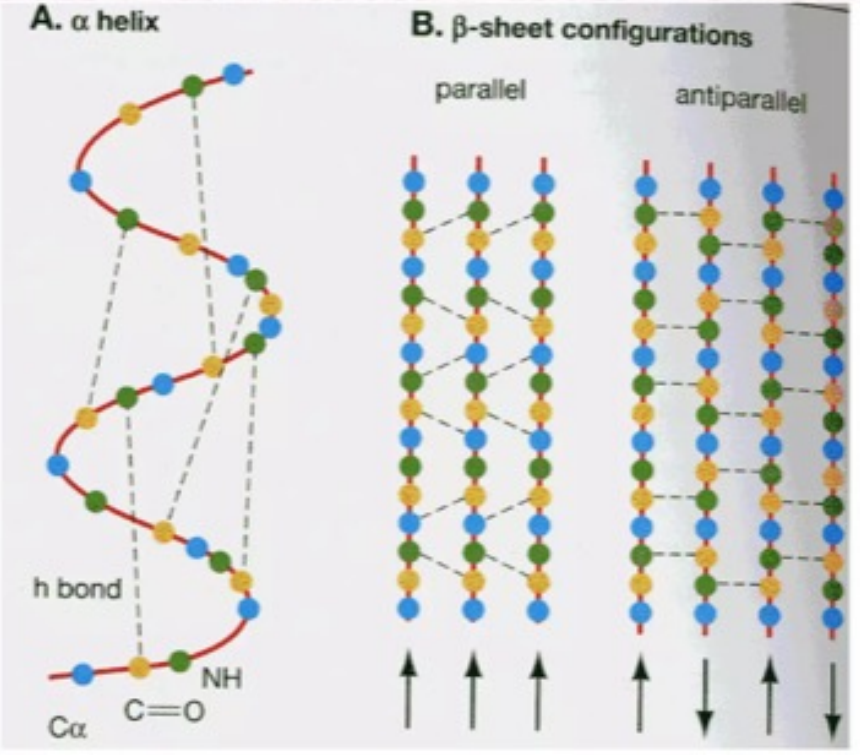
- Directions:
  - Same: Parallel Sheet
  - Opposite: Anti-parallel Sheet
  - Mixed: Mixed Sheet

- Pattern of hydrogen bond formation in parallel and anti-parallel sheets is different

β Sheet

- Slight counterclockwise rotation
- Alpha carbons (as well as R side groups) alternate above and below the sheet
- Prediction difficult, due to wide range of ϕ and ψ angles

Interactions in Helices and Sheets



Loop

- Regions between α helices and β sheets
- Various lengths and three-dimensional configurations
- Located on surface of the structure

Loop

- Hairpin loops: complete turn in the polypeptide chain, (anti-parallel β sheets)
- More variable sequence structure
- Tend to have charged and polar amino acids
- Frequently a component of active sites

Coil

- Region of secondary structure that is not a helix, sheet, or loop

Secondary Structure

6 Classes of Protein Structure

1) Class α: bundles of α helices connected by loops on surface of proteins

2) Class β: antiparallel β sheets, usually two sheets in close contact forming sandwich

3) Class α/β: mainly parallel β sheets with intervening α helices; may also have mixed β sheets (metabolic enzymes)

6 Classes of Protein Structure

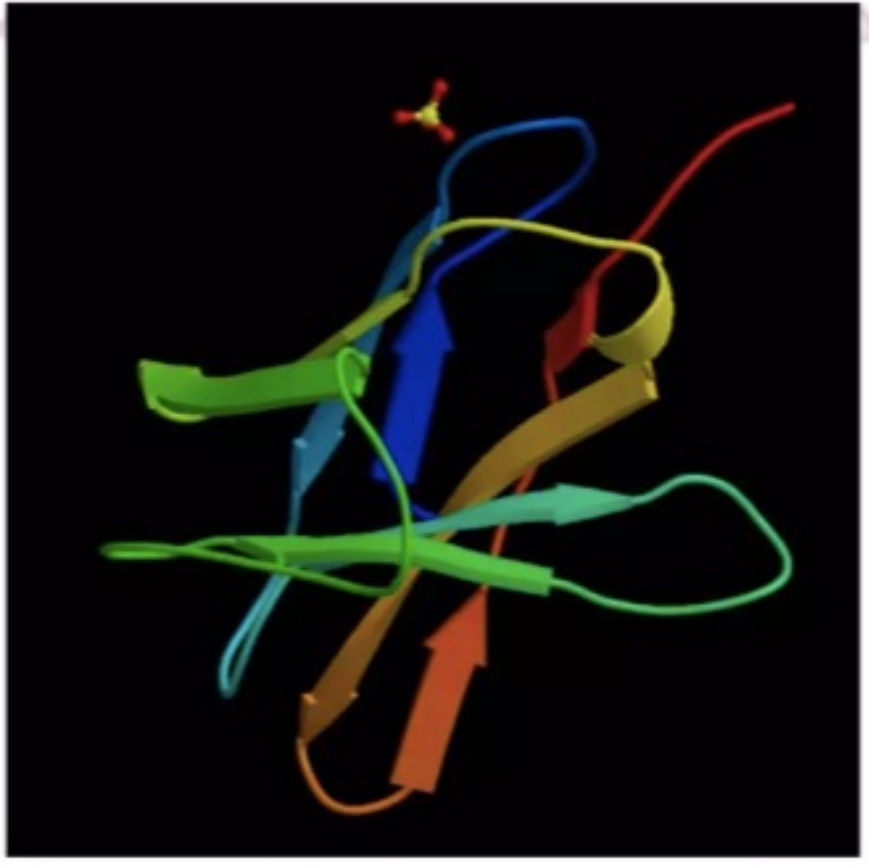
4) Class α+ β: mainly segregated α helices and antiparallel β sheets

5) Multidomain (α and β) proteins more than one of the above four domains

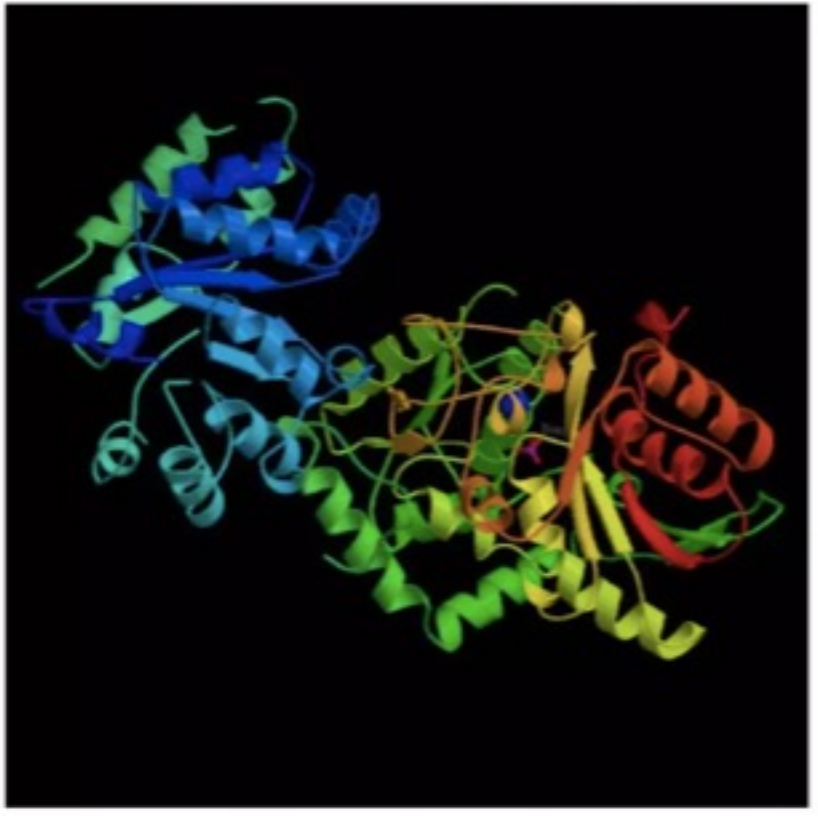
6) Membrane and cell-surface proteins and peptides excluding proteins of the immune system

α Class Protein (hemoglobin)

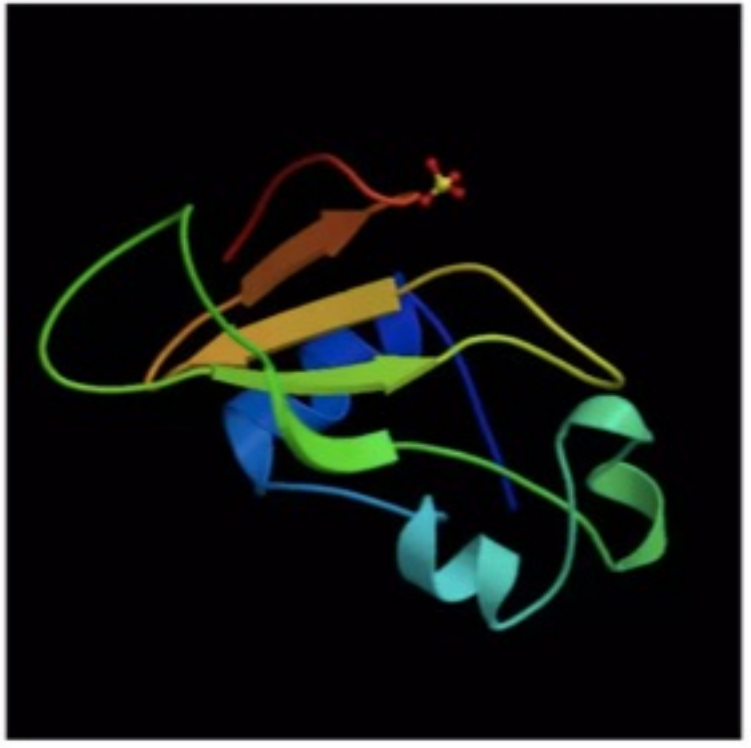
β Class Protein (T-Cell CD8)



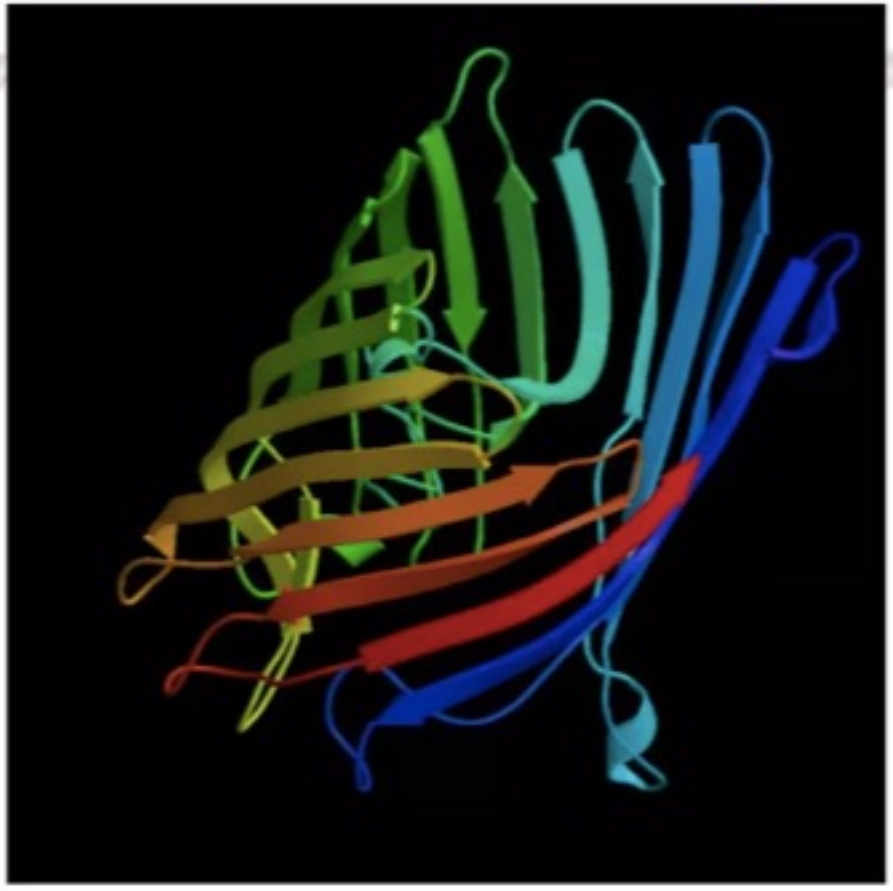
α/β Class Protein (tryptohan synthase)



α+β Class Protein (1RNB)



Membrane Protein (10PF)



Protein Structure Databases

- Databases of three dimensional structures of proteins, where structure has been solved using X-ray crystallography or nuclear magnetic resonance (NMR) techniques

- Protein Databases:
  - PDB
  - SCOP
  - Swiss-Prot
  - PIR

Protein Structure Databases

- Most extensive for 3-D structure is the Protein Data Bank (PDB)

Partial PDB File

```
ATOM      1  N   VAL A   1       6.452  16.459   4.843  7.00 47.38      3HHB 162
ATOM      2  CA  VAL A   1       7.060  17.792   4.760  6.00 48.47      3HHB 163
ATOM      3  C   VAL A   1       8.561  17.703   5.038  6.00 37.13      3HHB 164
ATOM      4  O   VAL A   1       8.992  17.182   6.072  8.00 36.25      3HHB 165
ATOM      5  CB  VAL A   1       6.342  18.738   5.727  6.00 55.13      3HHB 166
ATOM      6  CG1 VAL A   1       7.114  20.033   5.993  6.00 54.30      3HHB 167
ATOM      7  CG2 VAL A   1       4.924  19.032   5.232  6.00 64.75      3HHB 168
ATOM      8  N   LEU A   2       9.333  18.209   4.095  7.00 30.18      3HHB 169
ATOM      9  CA  LEU A   2      10.785  18.159   4.237  6.00 35.60      3HHB 170
ATOM     10  C   LEU A   2      11.247  19.305   5.133  6.00 35.47      3HHB 171
ATOM     11  O   LEU A   2      11.017  20.477   4.819  8.00 37.64      3HHB 172
ATOM     12  CB  LEU A   2      11.451  18.286   2.866  6.00 35.22      3HHB 173
ATOM     13  CG  LEU A   2      11.081  17.137   1.927  6.00 31.04      3HHB 174
ATOM     14  CD1 LEU A   2      11.766  17.306    .570  6.00 39.08      3HHB 175
ATOM     15  CD2 LEU A   2      11.427  15.778   2.539  6.00 38.96      3HHB 176
```

Description of PDB File

- second column: amino acid position in the polypeptide chain
- fourth column: current amino acid
- Columns 7, 8, and 9: x, y, and z coordinates (in angstroms)
- The 11th column: temperature factor -- can be used as a measurement of uncertainty

Protein Structure Classification Databases

- **Structural Classification of proteins (SCOP)**
- based on expert definition of structural similarities
- SCOP classifies by class, family, superfamily, and fold
- http://scop.mrc-lmb.cam.ac.uk/scop/

Protein Structure Classification Databases

- **Classification by class, architecture, topology, and homology (CATH)**
- Classifies proteins into hierarchical levels by class
- a/B and a+B are considered to be a single class
- http://www.biochem.ucl.ac.uk/bsm/cath/

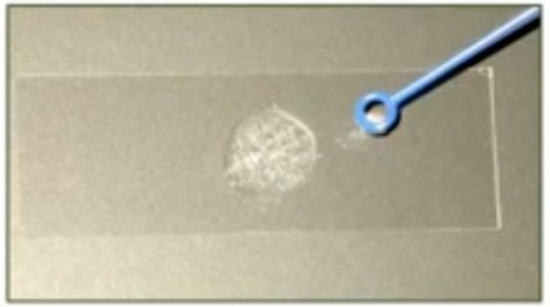
Protein Structure Classification Databases

- **Molecular Modeling Database (MMDB)**
- structures from PDB categorized into structurally related groups using the VAST
- looks for similar arrangements of secondary structural elements
- http://www.ncbi.nlm.nih.gov/Entrez

Contents

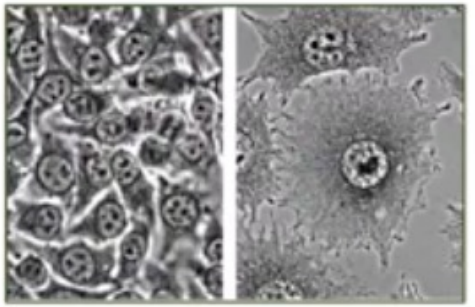
- Background
    a) **Chemical/Biochemical Reaction**
    b) **Molecular Recognition Process**
- Drug Discovery Process (A workflow)
- Ligand based / Structure based Drug Discovery
- Aims & Requirements of Docking
- Docking Algorithms for Ligand Sampling
    a) **Heuristic Approach**
    b) **Stochastic Approach**
    c) **Systematic Approach**
- Scoring Functions
    a) **Empirical Scoring Function**
    b) **Molecular Mechanics Force Fields**

Biochemical Reactions



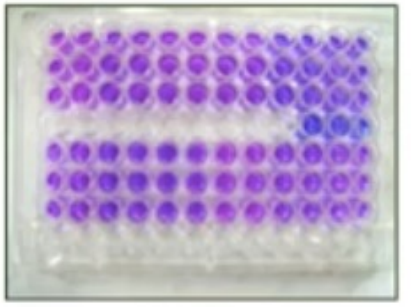
Catalase test

Catalase enzyme converts Hydrogen Peroxide to water and Oxygen



Target Id. By Image Analysis

Cells change morphology on addition of an inhibitor



Ligand Binding Assay

Colour of the reaction mixture change on ligand binding

Molecular Recognition

- Every cellular process happens due to interactions between molecules

- These interactions are governed by various intermolecular forces of attraction

Goal of docking:
Given structures of two biomolecules, determine if

- They interact favorably?
- If yes, then what is the orientation which maximizes interactions and minimizes energy

Interatomic Interactions

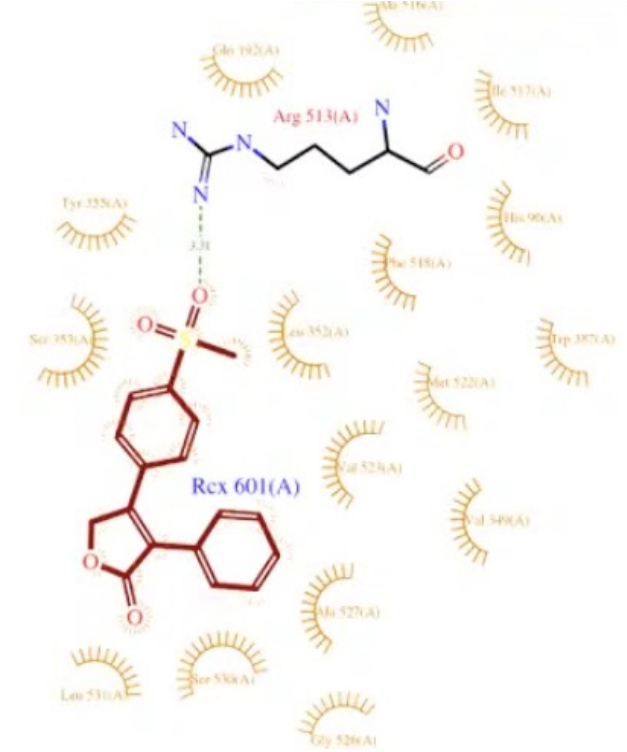
Enzyme-substrate complex is formed with geometric as well as electrostatic complementarity

• A large number of interactions contribute towards molecular recognition, including:

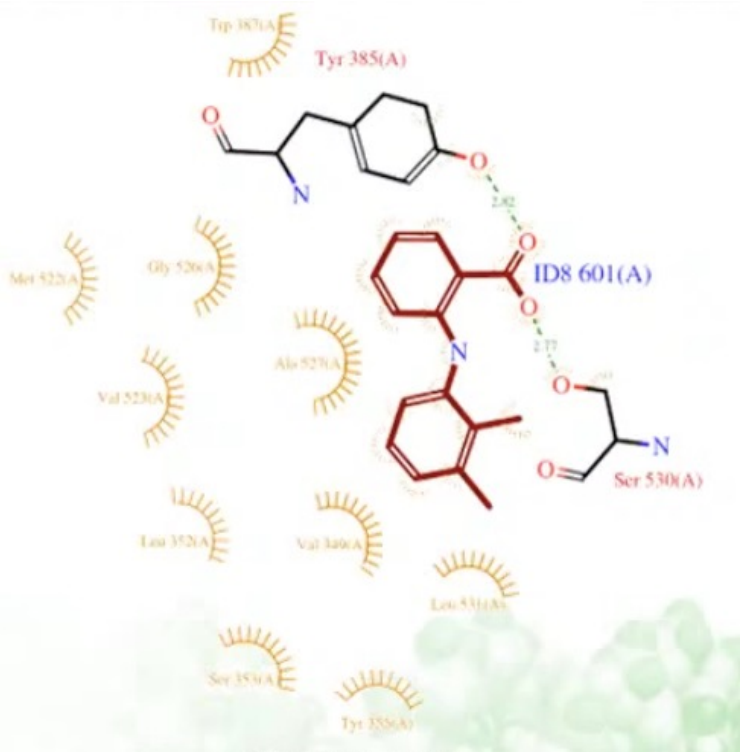
- Van der Waals, Hydrophobic, Hydrogen bond interactions, Ion - dipole, lon - ion, Covalent interactions

• Contribution of hydrogen bond interactions is significantly large in receptor-ligand interactions

H-Bond Contribution

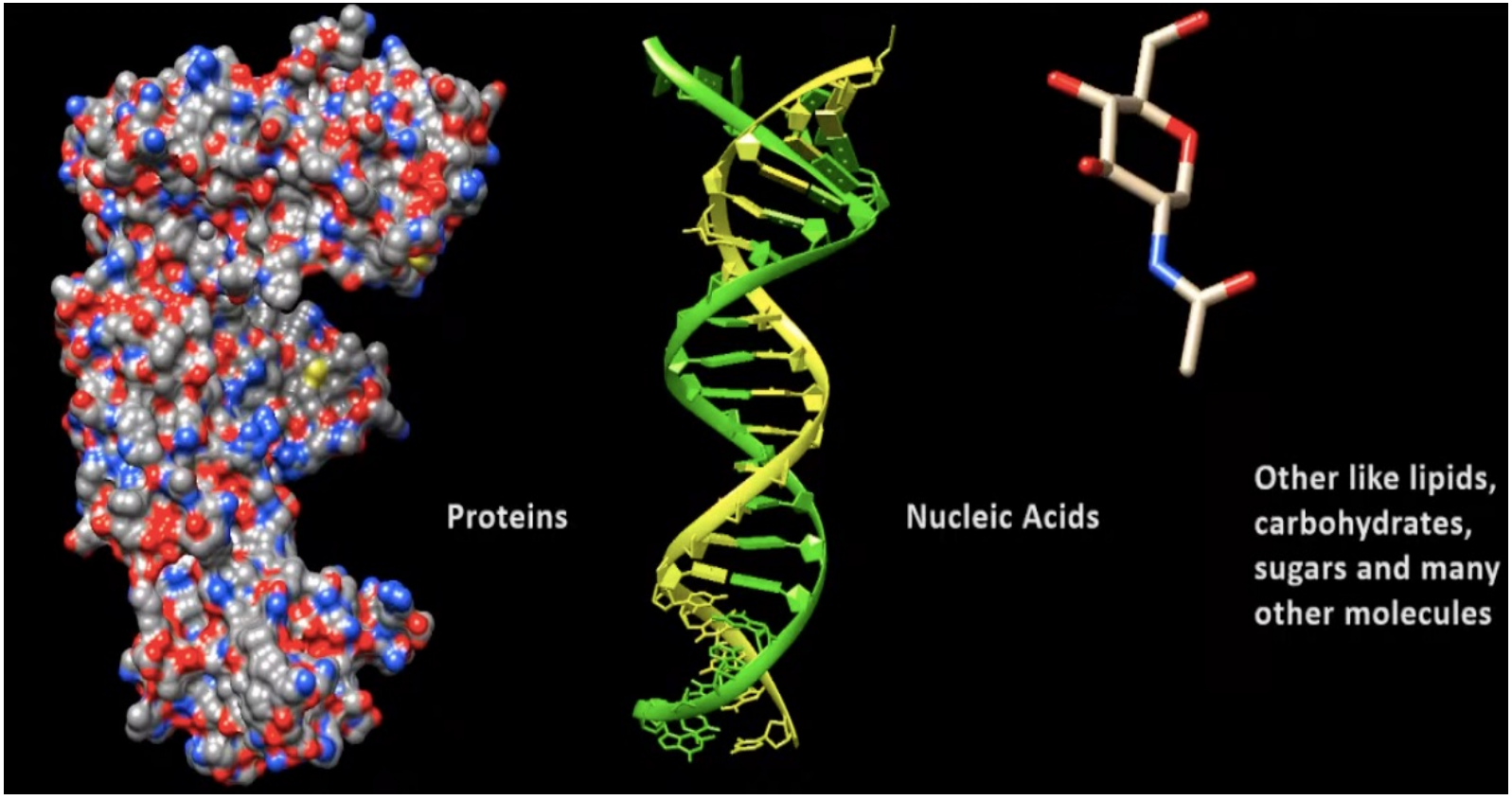


PDB ID: 5KIR
Vioxx (Rofecoxib) Bound to Cox-2
IC50: 3400nM

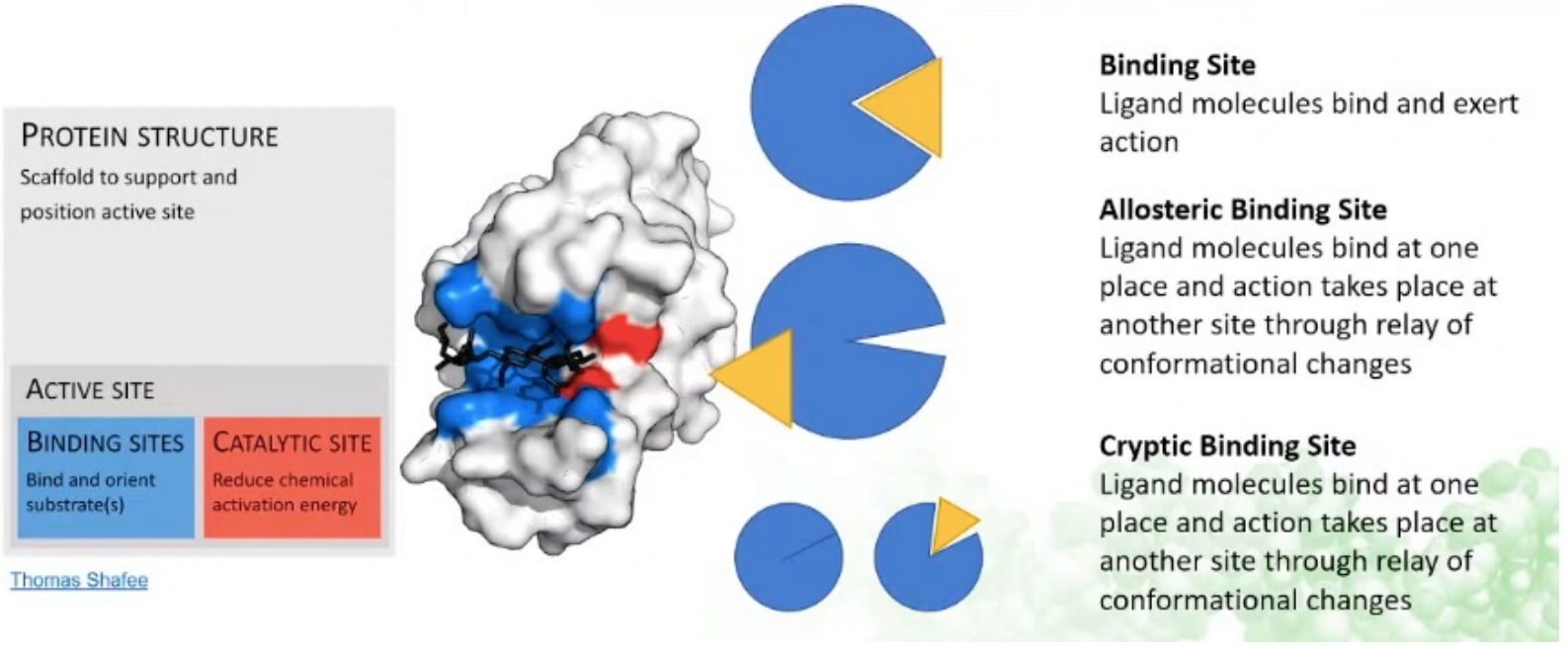


PDB ID: 5IKR
Mefenamic Acid Bound to Cox-2
IC50: 2900nM

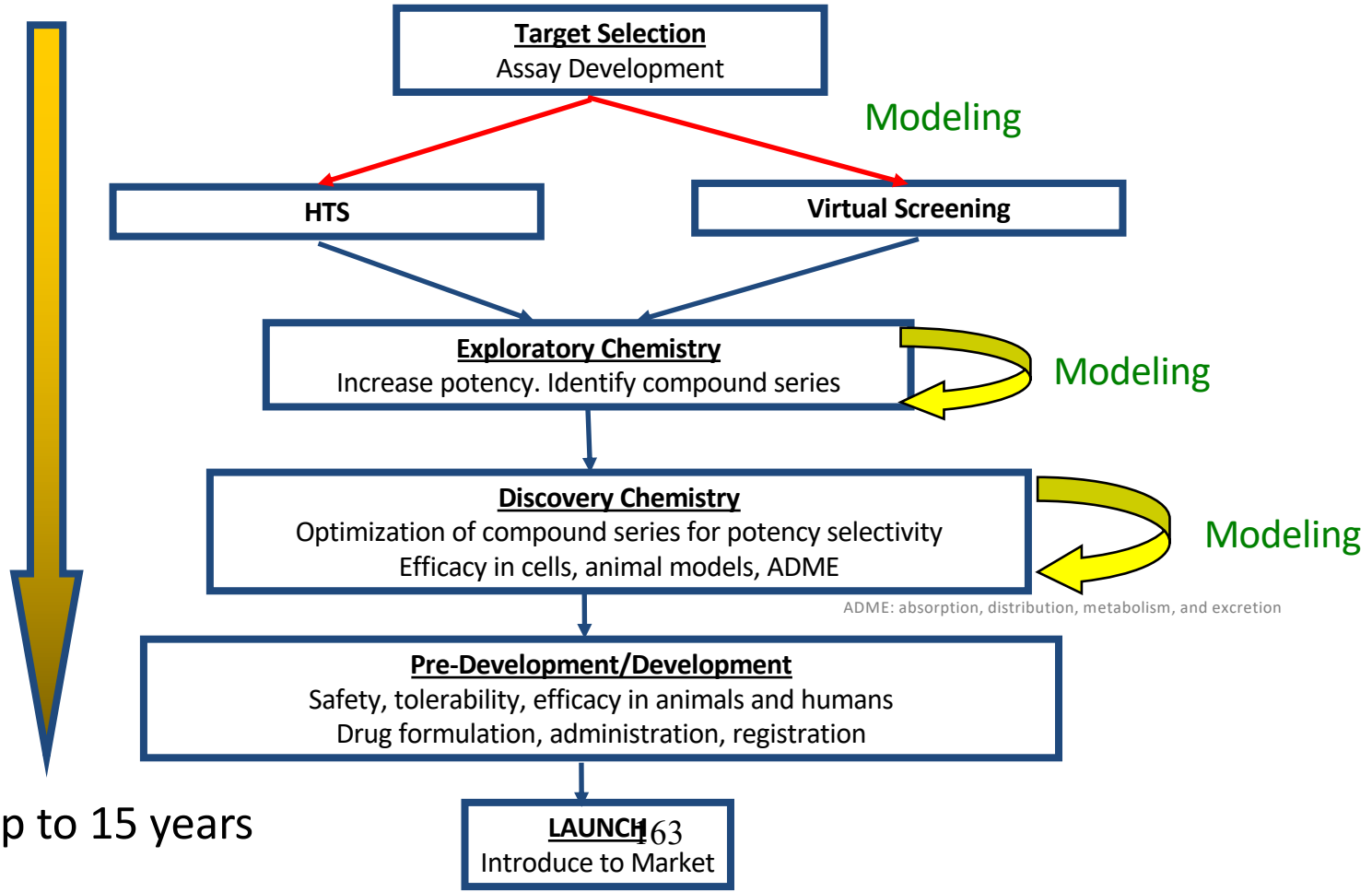
Biomolecular Structures



Binding Sites



Drug Discovery Process



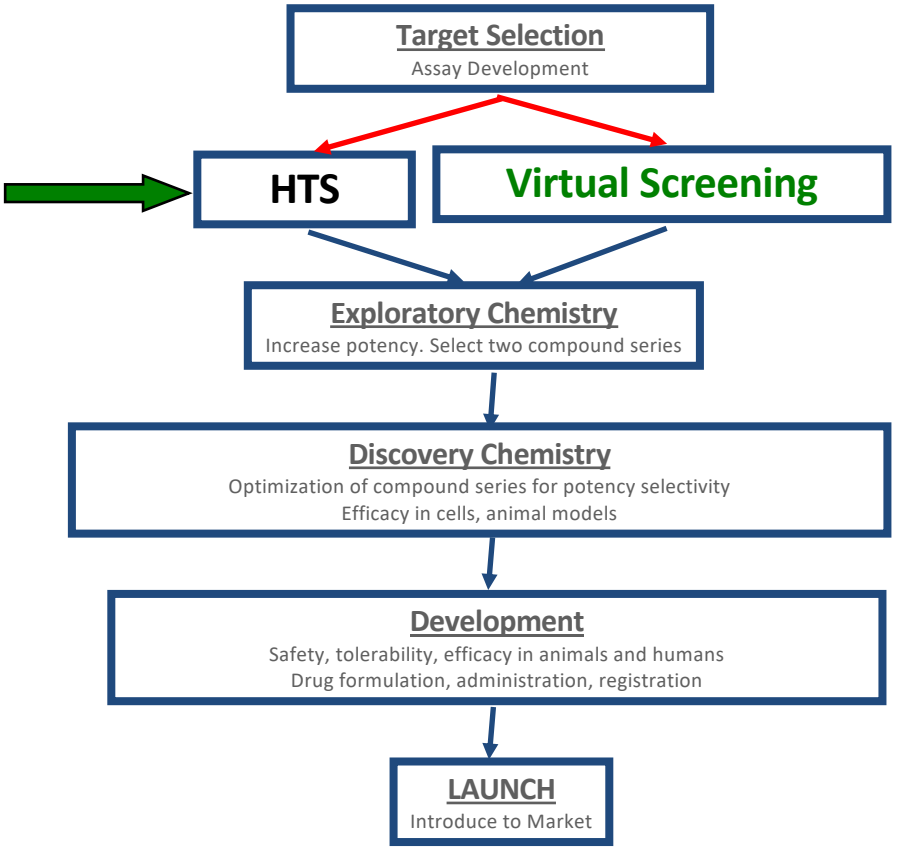
Drug Discovery Process: Database Filtering and Virtual Screening

Computational tools are used to weed-out compounds that may be

- Biologically inactive
- Toxic
- ADME
- **(Absorption, Distribution, Metabolism, and Excretion)**

Lipinski's rule is based on the statistical analysis of available drugs in the market

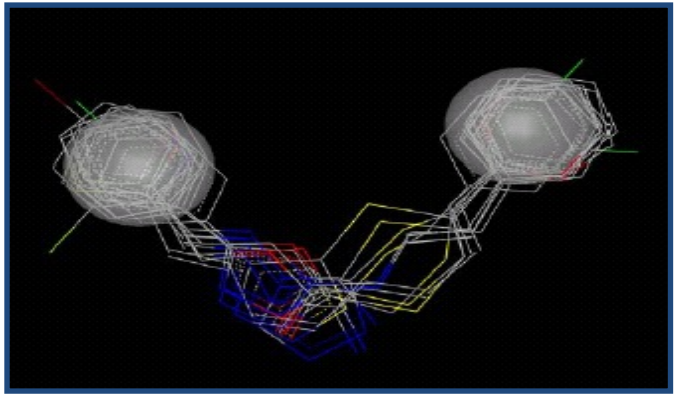
- Hydrogen bond donors <= 5
- Hydrogen bond acceptors <= 10
- Molecular weight < 500
- logP < 5
  - logP is a measure of hydrophobicity / hydrophilicity of the compound



Docking Approaches

2D Fragment-Based

- One or more structural features, either individually or in combination, of known active ligands are used to search database



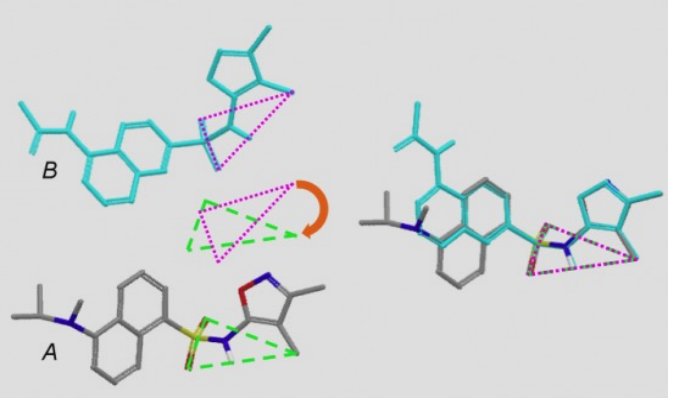
3D Ligand-Based
(Pharmacophore modeling)

- Requires a set of known active ligands
- Shape and electrostatics match

Flexible Ligand-Based Screening

(Shape Screening)

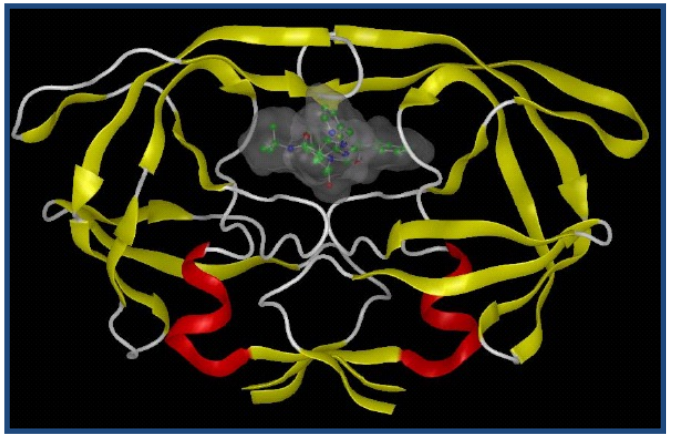
- Requires a set of known active ligands
- Shape-based flexible ligand superposition
- Ability to selectively identify actives over decoys within large database (gpu)



SimAB = VA∩B / VA∪B

Structure-Based (Docking)

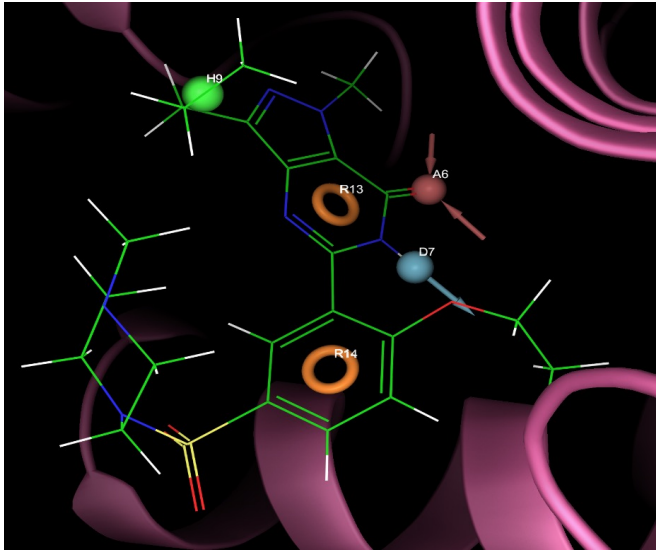
Requires a 3D structure of target protein
X-ray, NMR, Homology modeling
Requires knowledge of ligand binding site location



HIV protease

Structure-Based-Phase Screening

Phase uses energy terms of GlideXP docking to translate the binding interaction of the best docked poses.
It takes the cumulative energy contribution of all the constituting atoms of that site and selects the top ranked features for building the final pharmacophore.



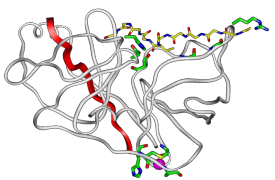
PDE

Docking Aims and Requirements

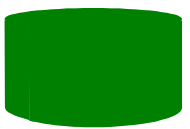
Aims

- Identify correct poses of ligands in the binding pocket
- Predict the affinity between the ligand and the protein.
  - Must be able to rank known ligands and screen out non-binders.

Requirements



Receptor



Compounds libraries
ZINC
PubChem
ChemDiv
NCI

Ligands

Computational Framework: Docking

Score the ligands

Protein-ligand Docking

- **Sampling Algorithms**
  - Heuristic (DOCK, FlexX)
  - Stochastic (GOLD)
  - Systematic (GLIDE)

- **Scoring Functions**
  - Empirical (FlexX, GOLD, GLIDE)
  - Molecular Mechanics (DOCK, GOLD, GLIDE)

Docking Algorithms Components 1
a) Sampling

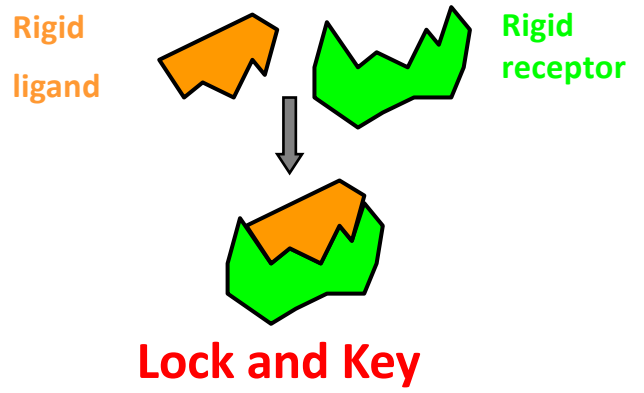
Increasing difficulty

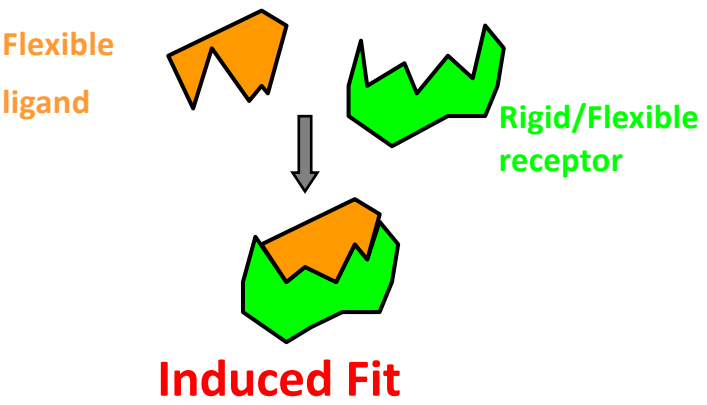
Rigid Receptor, Rigid Ligand

Rigid Receptor, Flexible Ligand

Flexible Receptor, Rigid Ligand

Flexible Receptor, Flexible Ligand





Sampling (conti..)

Complexity of the Docking Problem

Exhaustive search to find all possible binding modes

- Degrees of freedom: 3 translations, 3 rotations, Internal conformational degree of freedom

Estimation of search space

- Typical binding pocket: Active site space: 10^3 Å^3
- Typical ligand: 4 rotational bonds
- Sampling space
  - Angle sampling: 10^0
  - Translational: 0.5 Å grid
- Compute speed: 10 conformations per second

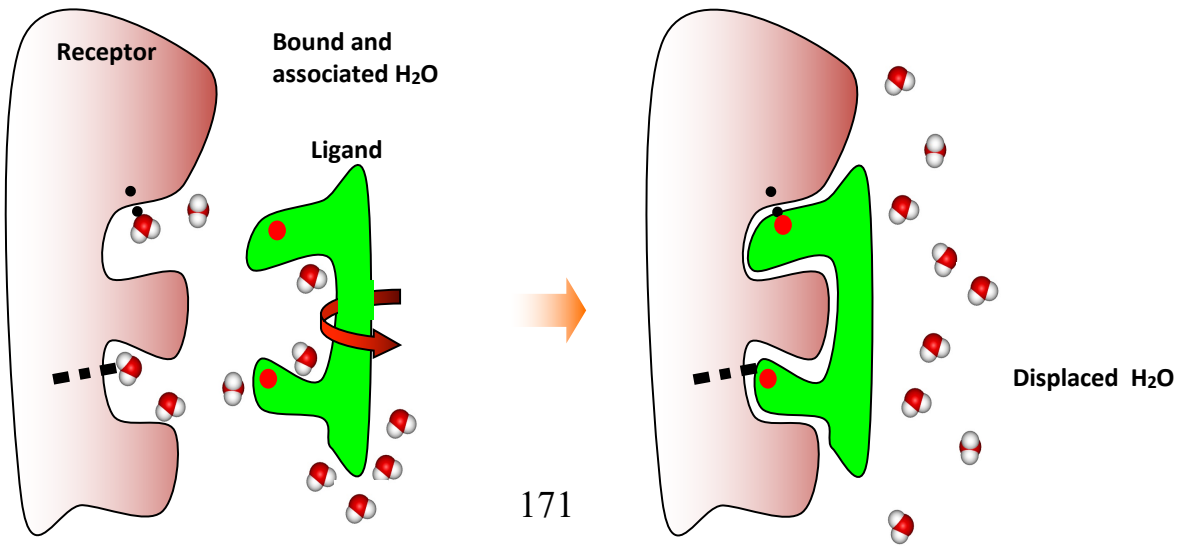
Size of search space

$4x10^8$ orientations / ligand

(and we have not talked about the receptor yet…)

Docking Algorithms Component 2
Scoring

- The ideal scoring function for protein-ligand interaction must take into account and accurately calculate:
  - Desolvation energies of receptor and ligand
  - Protein-ligand interaction energy
  - Strain energies of receptor and ligand after complex formation
  - Solvation energy of protein-ligand complex
  - Change in entropy of ligand, receptor, and solvent



Protein-Ligand Docking

- **Sampling Algorithms**
  - Heuristic (DOCK, FlexX)
  - Stochastic (GOLD)
  - Systematic (GLIDE)

- **Scoring Functions**
  - Empirical (FlexX, GOLD, GLIDE)
  - Molecular Mechanics (DOCK, GOLD, GLIDE)

DOCK Program for Rigid Docking:
Rigid Receptor and Rigid Ligand

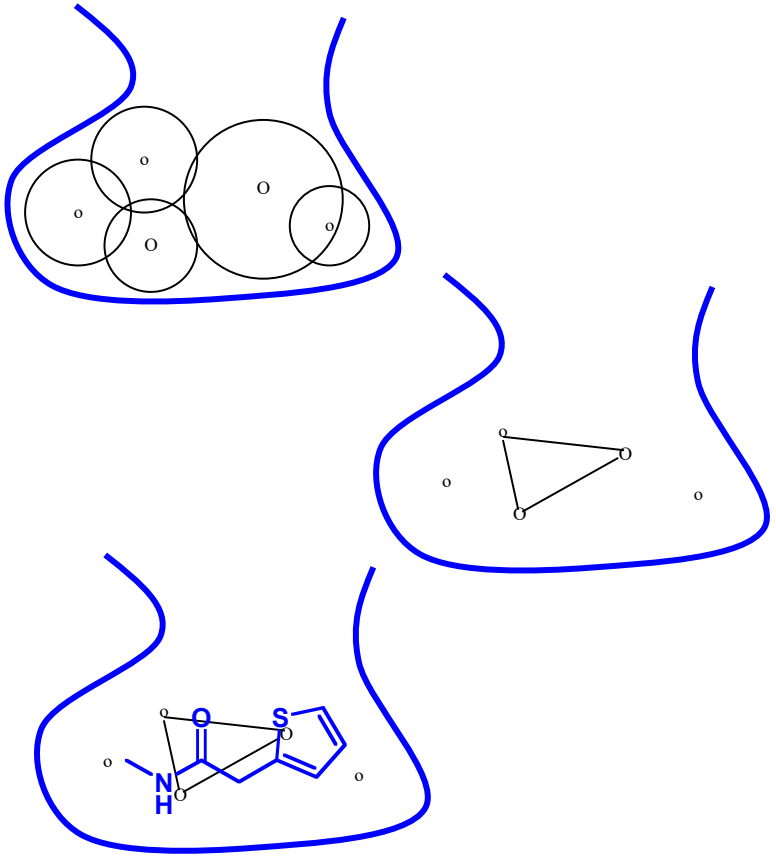
DOCK works in 4 steps:

Step 1 Generate molecular surface for receptor

Step 2 Generate spheres to fill the active site of the receptor: The spheres become potential locations for ligand atoms

Step 3 Matching: Sphere centers are then matched to the ligand atoms, to determine possible orientations for the ligand

Step 4 Scoring: Find the top scoring orientation



Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. L. & Ferrin, T. E. (1982). A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* **161**, 269–288.

DOCK Program for Rigid Docking:

Speed and Accuracy

- Advantage
  - Speed: Computationally efficient

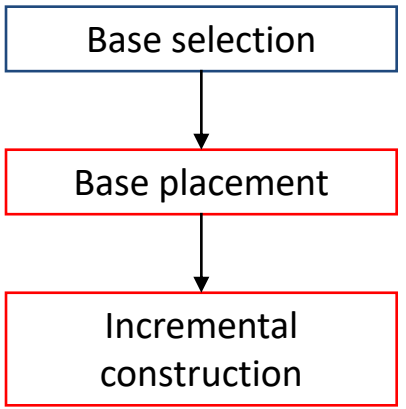
- Disadvantage
  - Accuracy: Minimum energy ligand conformation often does not correspond to the bound conformation

Docking by Incremental Construction (FlexX):

Rigid Receptor and Flexible Ligand

- FlexX incorporates ligand flexibility using an incremental construction algorithm
  - An alternative to exhaustive exploration of ligand conformations

Components of the FlexX algorithm



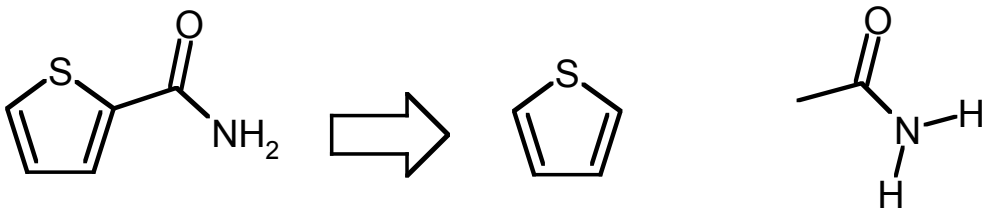
A Fast Flexible Docking Method using an Incremental Construction Algorithm

Matthias Rarey[1]*, Bernd Kramer[1], Thomas Lengauer[1] and Gerhard Klebe[2]

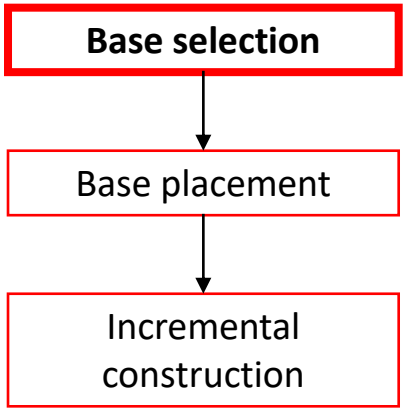
J. Mol. Biol. (1996) **261**, 470–489

FlexX Algorithm (1)

- Split ligand into base fragments



- Base fragment should form directional interactions with the receptor
- Large, flexible base fragment increases computational complexity
- FlexX is sensitive to the selection of the base fragment
- Criteria for base fragment selection:
  - Maximize the number of potential interaction groups
  - Minimize the number of alternative conformations

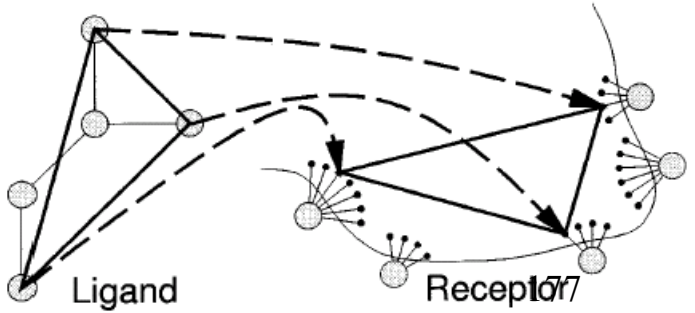


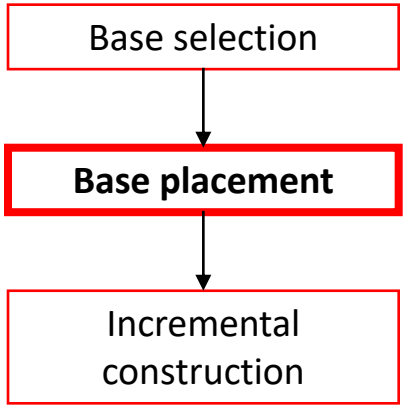
FlexX Algorithm (2)

- Base fragments are positioned in the active site to form maximum number of favorable interactions with the receptor
  - Receptor is approximated as a discrete set of interaction points

- Three interaction points from base fragment are mapped onto three interaction points of the receptor.

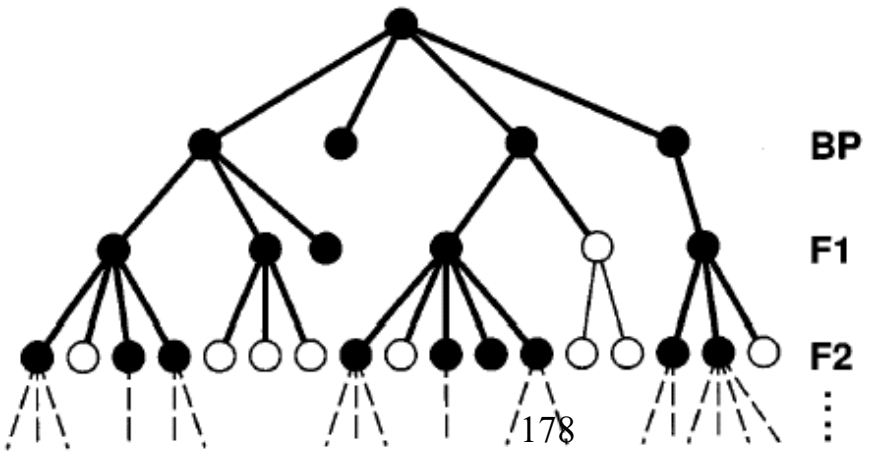
- Rigid body superposition of the interaction centers

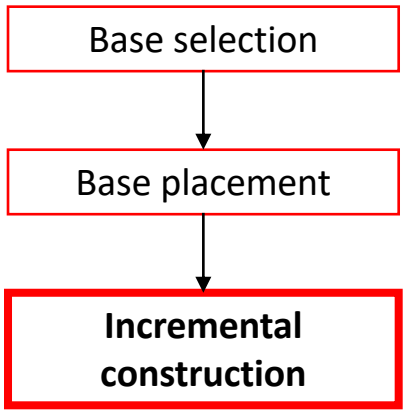




FlexX Algorithm (3)

- All possible placements of the base fragments (BP)
- Level 2: Add next fragment F1. Priority given to the fragment capable of forming directional interaction to the receptor
- Continue adding remaining fragments, F2, F3 …





Flex: Speed and Accuracy

- Advantage
  - Speed: Incremental construction strategy provides an efficient alternative to exhaustive exploration of the ligand conformational space
- Disadvantage
  - Accuracy: Too sensitive to the selection of base fragment

Protein-ligand Docking

- **Sampling Algorithms**
  - Heuristic (DOCK, FlexX)
  - Stochastic (GOLD)
  - Systematic (GLIDE)

- **Scoring Functions**
  - Empirical (FlexX, GOLD, GLIDE)
  - Molecular Mechanics (DOCK, GOLD, GLIDE)

Genetic Optimization for Ligand Docking (GOLD):

Partially Flexible Receptor and Flexible Ligand

- Genetic Algorithm
  - Based on the process of Darwinian Evolution -- **survival of the fittest**

  - Three components:
    - Representation of the problem as a chromosome
    - Scoring function to determine fitness
    - Genetic operator

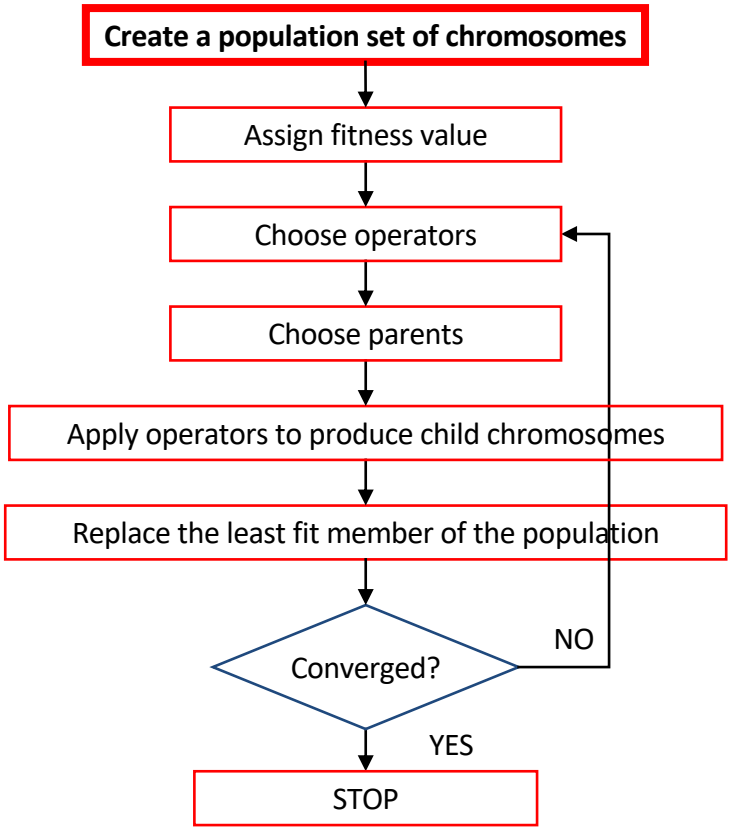
Development and Validation of a Genetic Algorithm for Flexible Docking

Gareth Jones[1*], Peter Willett[1], Robert C. Glen[2], Andrew R. Leach[3] and Robin Taylor[4]

J. Mol. Biol. (1997) **267**, 727–748

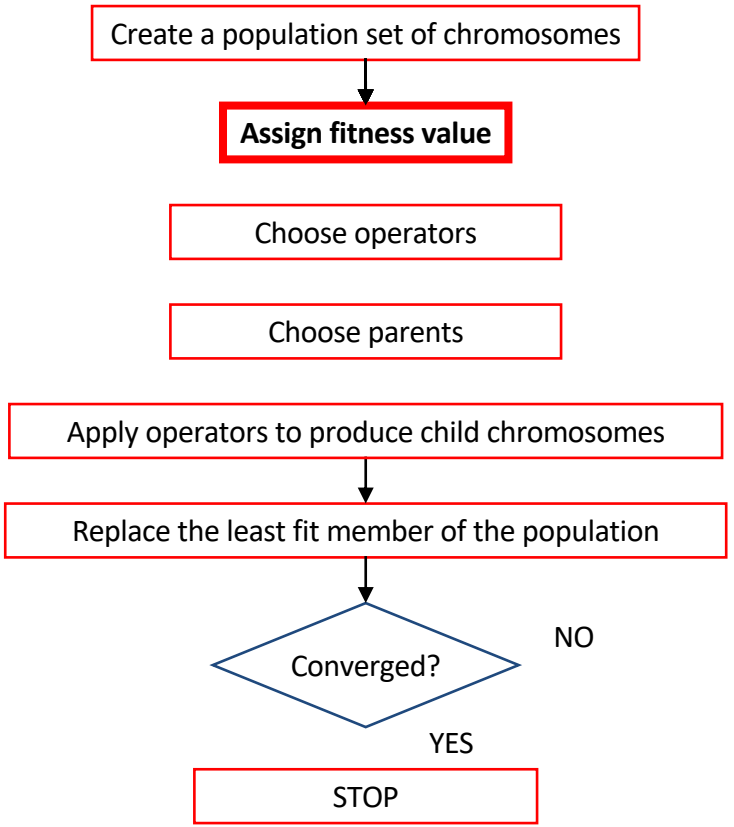
GOLD Algorithm (1)

- Each “chromosome” codes for ligand conformation and orientation of the ligand
- Initial population set is generated randomly



GOLD Algorithm (2)

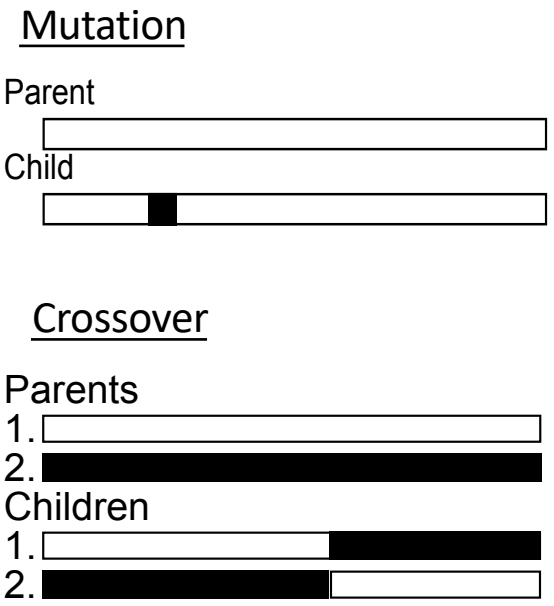
- Fitness function is a combination of ligand-protein hydrogen bond and van der waals interaction energies
- Fitness function also contains of a term for ligand strain energy

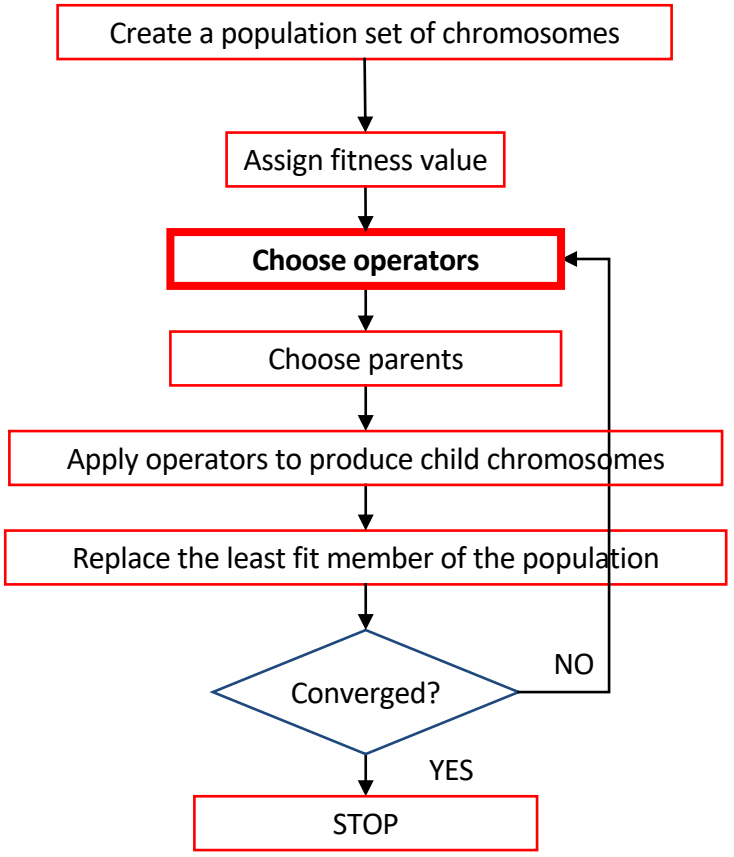


GOLD Algorithm (3)

Crossover and Mutation operators

- Crossover requires two parents and generates two children
- Mutation requires one parent and produces one child

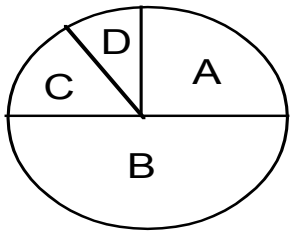


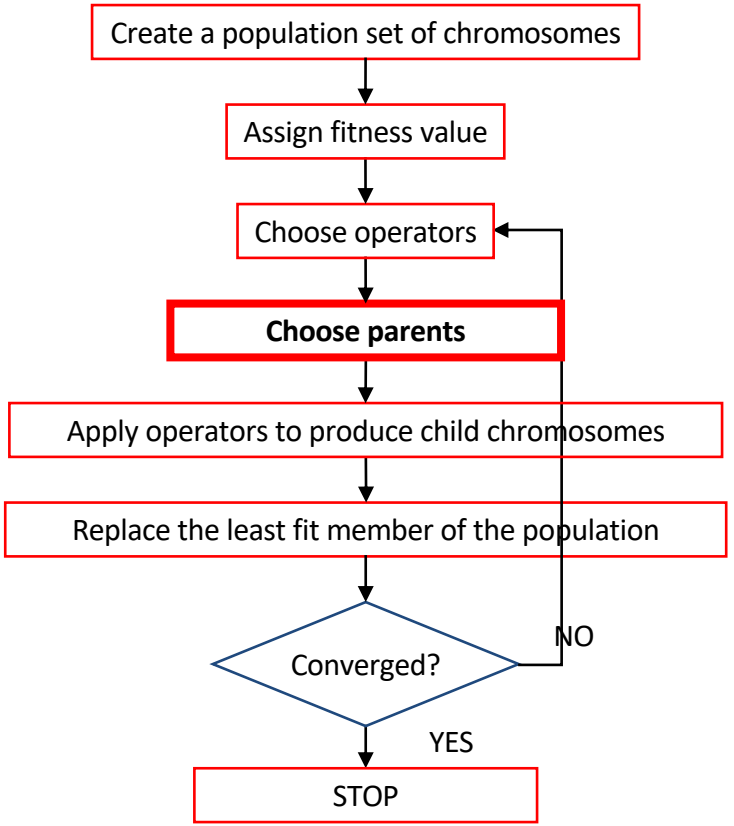


GOLD Algorithm (4)

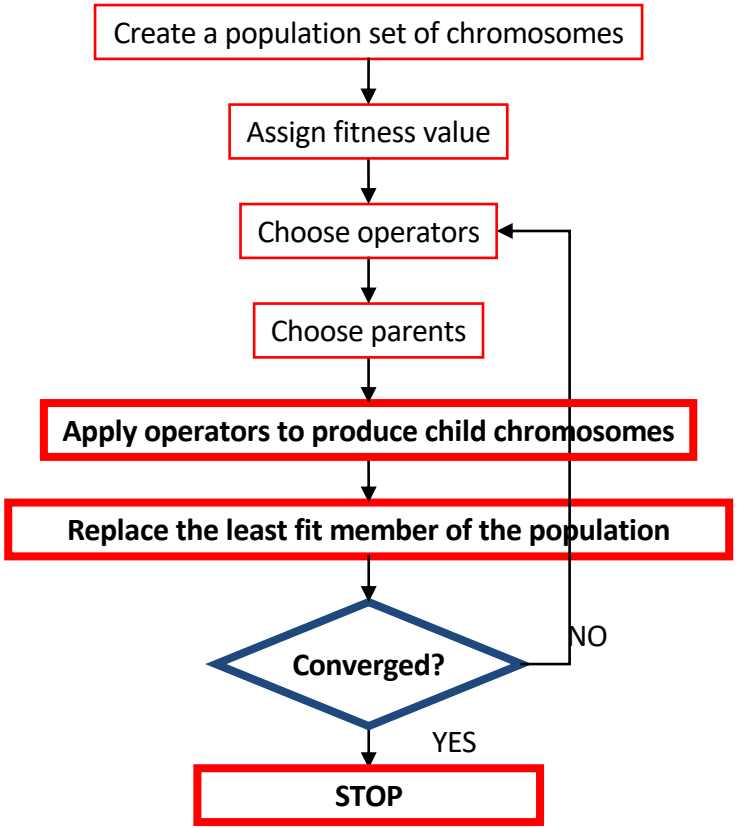
Selection is biased by fitness score

Population	Fitness Score
A	**3**
B	**6**
C	**2**
D	**1**





GOLD Algorithm (5)



GOLD: Speed and Accuracy

- Advantage
  - Speed: Directly proportional to the sampling. Multiple runs with different starting population sets are required

- Disadvantage
  - Accuracy: Due to the stochastic nature of the algorithm, it is easy to miss the bioactive ligand conformation

Protein-Ligand Docking

- **Sampling Algorithms**
  - Heuristic (DOCK, FlexX)
  - Stochastic (GOLD)
  - Systematic (GLIDE)

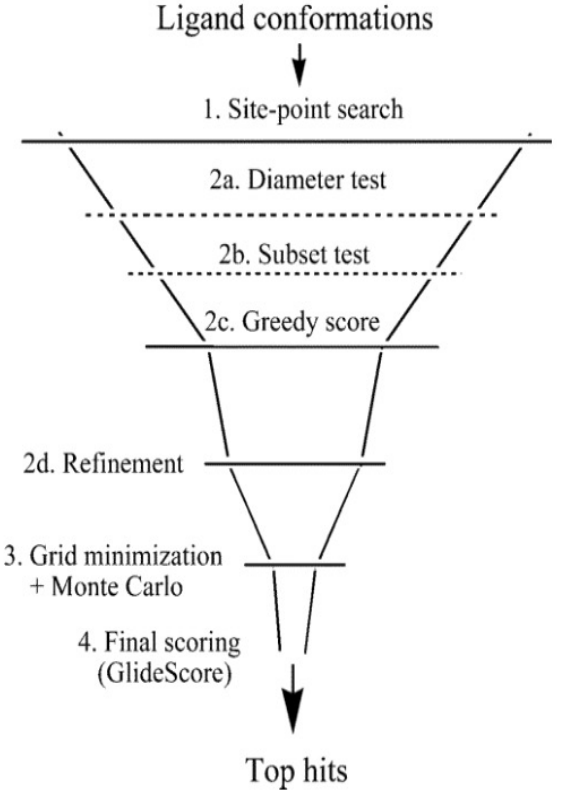
- **Scoring Functions**
  - Empirical (FlexX, GOLD, GLIDE)
  - Molecular Mechanics (DOCK, GOLD, GLIDE)

Grid-Based Ligand Docking with Energetics (**GLIDE**)

Rigid Receptor and Flexible Ligand

Key features of GLIDE

- Complete systematic search of the conformational, orientational, and positional space of the docked ligand
- Hierarchical filtering
- Shape and properties of receptor are represented on a grid
- Exhaustive enumeration of ligand torsions generated

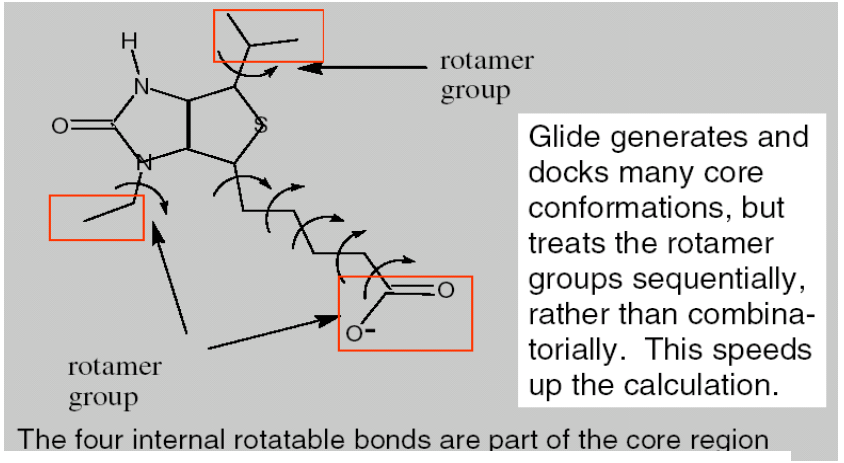


Docking using GLIDE (2)

Step 1**: Ligand Conformation Generation**

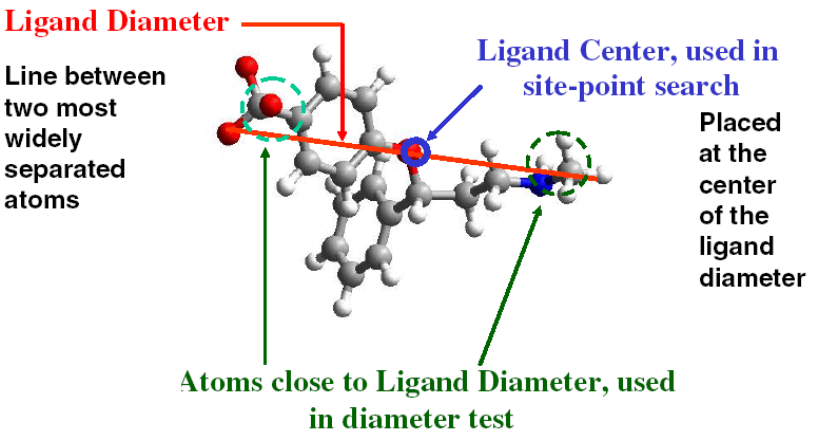
Explore entire phase space of ligand

Reduces the region of phase space which are energetically not favorable



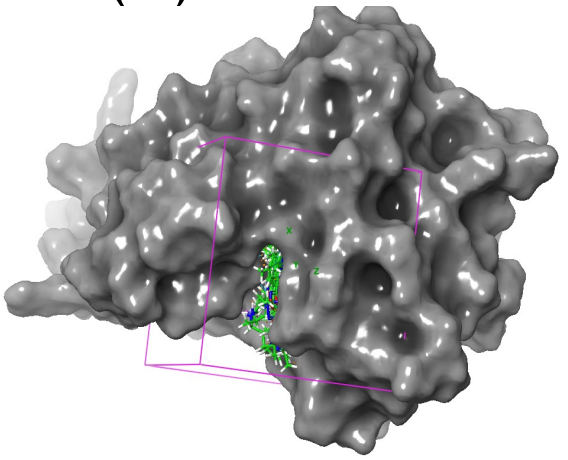
Step 2: **Ligand Placement**

Position and orientation relative to the receptor, core conformation, and rotamer-group conformations.



Docking using GLIDE (3)

Step 3: **Site Point Search**



Selection of “site points” on an equally spaced 2 Å grid that covers the active site region

Step 4: Rough Scoring (OPLS-AA forcefield scoring)

Step 5: Energy Minimization (Monte Carlo minimization)

Step 6: Final Scoring (modified Chemscore scoring + MM)

Protein-ligand Docking

- **Sampling Algorithms**
  - Heuristic (DOCK, FlexX)
  - Stochastic (GOLD)
  - Systematic (GLIDE)

- **Scoring Functions**
  - Empirical (FlexX, GOLD, GLIDE)
  - Molecular Mechanics (DOCK, GOLD, GLIDE)

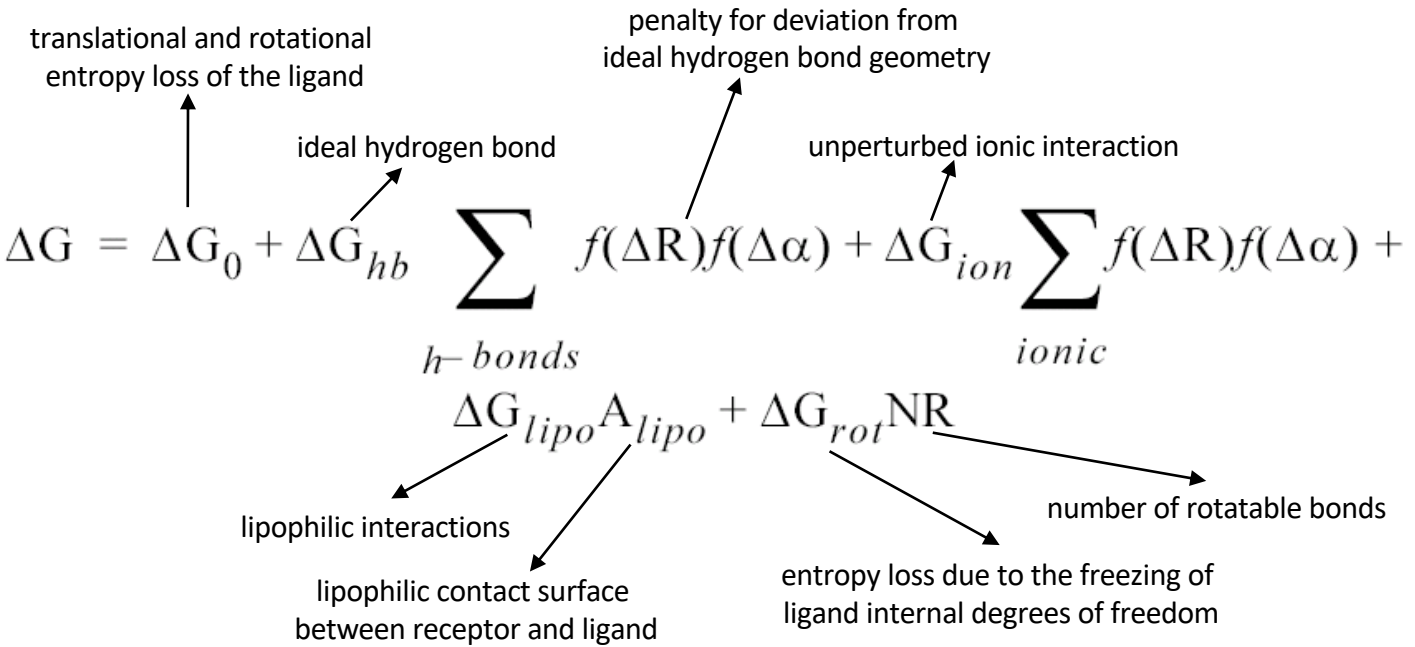
Empirical Scoring Function

- Ligand-protein binding affinity is approximated as a sum of a series of terms
- Parameters are derived using experimental binding energies from a training set of known protein-ligand complex structures
- Example: Bohm empirical scoring function

$$\Delta G = \Delta G_0 + \Delta G_{hb} \sum_{h-bonds} f(\Delta R) f(\Delta \alpha) + \Delta G_{ion} \sum_{ionic} f(\Delta R) f(\Delta \alpha) + \Delta G_{lipo} A_{lipo} + \Delta G_{rot} NR$$

H. J Bohm. J Comput Aided Mol Des. 8:243-256, 1994

Bohm Empirical Scoring Function



H. J Bohm. J Comput Aided Mol Des. 8:243-256, 1994

Application of Empirical Scoring Functions

Bohm Empirical Scoring Function

$$\Delta G = \Delta G_0 + \Delta G_{hb} \sum_{h-bonds} f(\Delta R) f(\Delta \alpha) + \Delta G_{ion} \sum_{ionic} f(\Delta R) f(\Delta \alpha) + \Delta G_{lipo} A_{lipo} + \Delta G_{rot} NR$$

- FlexX extends Bohm scoring function by splitting the lipophilic interaction term into two groups: (1) aromatic and (2) others
- GOLD uses empirical hydrogen bonding interaction
- GLIDE uses Chemscore; an empirical scoring scheme similar to Bohm scoring function
- **Advantage:** Computationally efficient. Based on known physical principles
- **Disadvantage:** Not always transferable to protein-ligand complexes that are very different from the training set

Protein-ligand Docking

- **Sampling Algorithms**
  - Heuristic (DOCK, FlexX)
  - Stochastic (GOLD)
  - Systematic (GLIDE)

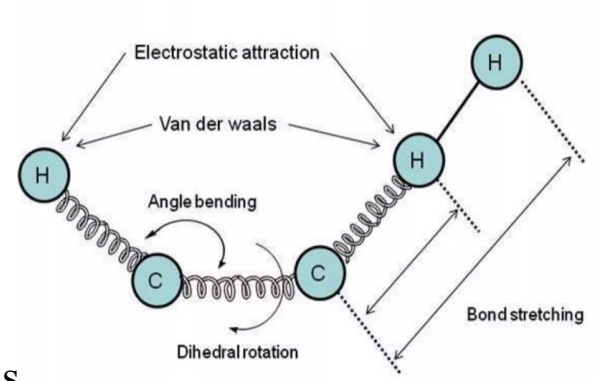
- **Scoring Functions**
  - Empirical (FlexX, GOLD, GLIDE)
  - Molecular Mechanics (DOCK, GOLD, GLIDE)

Molecular Mechanics Force Fields

Molecular mechanics force fields are used to estimate the enthalpic contributions to the free energy.

Entropic contributions are calculated separately

$$\Delta G_{bind} = E_{MM} - T\Delta S_{solute} + \Delta G_{solvent}.$$



Functional form of molecular mechanics force fields

Penalty for deviation from ideal bond length, bond angles, and dihedral angles

$$E_{MM} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\vartheta(\vartheta - \vartheta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]$$

$$+ \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right].$$

van der waals interactions

electrostatics interactions

Application of Molecular Mechanics Force Fields

$$E_{MM} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\vartheta(\vartheta - \vartheta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]$$
$$+ \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right].$$

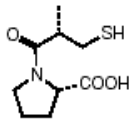
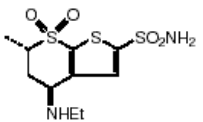
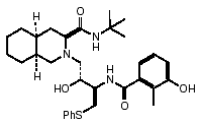
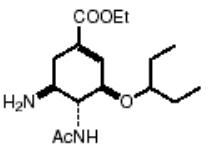
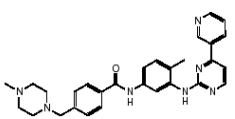
DOCK scoring function uses van der Waals and Coulomb interaction terms.

GOLD uses van der Waals term for receptor-ligand interaction and ligand strain energy.

GLIDE uses van der Waals and Coulomb interactions terms as well as ligand strain energy.

- **Advantage:** Easy and fast to calculate.
- **Disadvantage:** Entropic and solvent contributions to the free energy are difficult to determine accurately.

Example of drugs derived from structure-based approaches

Name			target	disease	year	company
Capoten	*Captopril*		*ACE*	*Hypertension*	*1981*	*Bristol-Myers Squibb*
Trusopt	*Dorzolamide*		*Carbonic anhydrase*	*Glaucoma*	*1995*	*Merck*
Viracept	*Nelfinavir*		*HIV protease*	*HIV/ AIDS*	*1999*	*Agouron (Pfizer) and Lilly*
Tamiflu	*Oseltamivir*		*Neuraminidase*	*Influenza*	*1999*	*Gilead and Roche*
Gleevec	*Imatinib*		*BCR- Abl*	*Chronic myelogenous leukaemia*	*2001*	*Novartis*

Take Home Points

- *Molecular modeling is important for structure based drug discovery*
- *Docking is a useful tool for lead identification and lead optimization*
- *Various methods for protein-ligand docking exist, but none of these can be considered superior to the rest*
- *Performance of protein-ligand docking is limited by the accuracy of current scoring functions*

Thank you