



# Bioinformatica II

Alberto Pallavicini



# Sequence Comparisons

- Homology searches
  - Usually 'one-against-one': *BLAST, FASTA*
  - Allows for comparison of individual sequences against databases comprised of individual sequences
- Profile searches
  - Uses collective characteristics of a family of proteins
  - Search can be 'one-against-many': *Pfam, CDD*  
or 'many-against-one': *PSI-BLAST, DELTA-BLAST*



# Profiles, Patterns, Motifs, and Domains

## Profiles

- Numerical representations of multiple sequence alignments
- Depend upon *patterns* or *motifs* containing conserved residues
- Represent the common characteristics of a protein family
- Can find similarities between sequences with little or no sequence identity
- Allow for the analysis of distantly related proteins

# Profile Construction

APHIIIVATPG  
 GCEIVIAATPG  
 GVEICIAATPG  
 GVDILIGTTG  
 RPHIIIVATPG  
 KPHIIIAATPG  
 KVQLIIAATPG  
 RPDIVIAATPG  
 APHIIIVGTPG  
 APHIIIVGTPG  
 GCHVVIAATPG  
 NQDIVVATTG

- Which residues are seen at each position?
- What is the frequency of observed residues?
- Which positions are conserved?
- Where can gaps be introduced?

Position-Specific Scoring Table

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	0	13	0	0	-12	13	0	0	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	0	7	0	0	-11	13	11	-3	0	-16	-11	-2	89	17	17	24	22	9	-50	-48	12
G	70	00	20	70	30	-80	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

## Patterns

*Phe  
or Tyr*

*Cys*

*not Val  
or Ala*

*three  
His*

$[FY] - x - C - x(2) - \{VA\} - x - H(3)$

*any  
amino  
acid*

*any two  
amino  
acids*

*any  
amino  
acid*





# PFAM

- Collection of multiple alignments of protein domains and conserved protein regions that probably have structural, functional, or evolutionary importance
- Each Pfam entry contains:
  - Multiple sequence alignment of family members
  - Protein domain architectures
  - Species distribution of family members
  - Information on known protein structures
  - Links to other protein family databases



## Pfam 32.0 (September 2018, 17929 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [Less...](#)

Proteins are generally composed of one or more functional regions, commonly termed **domains**. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

Pfam also generates higher-level groupings of related entries, known as **clans**. A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-HMM.

The data presented for each entry is based on the [UniProt Reference Proteomes](#)<sup>1</sup> but information on individual UniProtKB sequences can still be found by entering the protein accession. Pfam *full* alignments are available from searching a variety of databases, either to provide different accessions (e.g. all UniProt and NCBI GI) or different levels of redundancy.



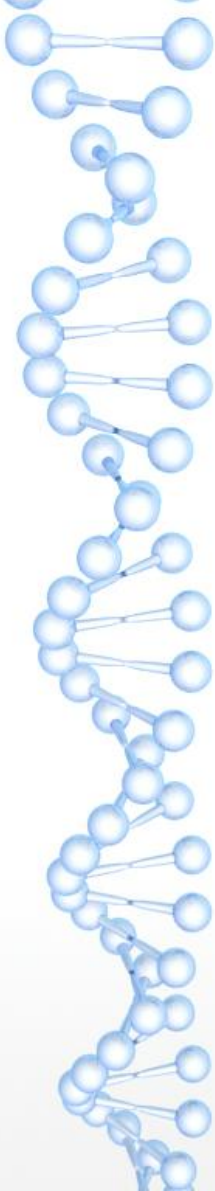
A profile HMM is a variant of an HMM relating specifically to biology  
They capitalise on the fact that certain positions in a sequence align






# PFAM A

- Based on *curated* multiple alignments of known members of a protein family ('seed alignment')
  - *Pfam definition of 'family': a collection of related protein regions*
  - *Based on reference proteomes (UniProtKB)*
- HMMER used to find all detectable protein sequences belonging to the family
- New 'true members' of the family are then used to generate the 'full alignment' for the protein family
- Given the method used to construct the alignments, hits are highly likely to be true positives




EMBL-EBI  [HOME](#) | [SEARCH](#) | [BROWSE ABOUT](#) | [FTP](#) | [HELP](#)

**Pfam**  
keyword search [Go](#)

**Pfam 29.0 (December 2015, 16295 entries)**

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

---

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
<a href="#">SEQUENCE SEARCH</a>	Analyze your protein sequence for Pfam matches 
<a href="#">VIEW A PFAM ENTRY</a>	View Pfam annotation and alignments
<a href="#">VIEW A CLAN</a>	See groups of related entries
<a href="#">VIEW A SEQUENCE</a>	Look at the domain organisation of a protein sequence
<a href="#">VIEW A STRUCTURE</a>	Find the domains on a PDB structure
<a href="#">KEYWORD SEARCH</a>	Query Pfam by keywords
<a href="#">JUMP TO</a>	<input type="text"/> <a href="#">Go</a> <a href="#">Example</a> Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc. Or view the <a href="#">help</a> pages for more information



# CDD Conserved Database Domain

- Identify conserved domains in a protein sequence
- Incorporates three-dimensional structural information to define domain boundaries and refine alignments
- Source data derived from:
  - Pfam A
  - Simple Modular Architecture Research Tool (SMART)
  - COG (orthologous prokaryotic protein families)
  - PRK ('protein clusters' of related protein RefSeq entries)
  - TIGRFAM



<http://ncbi.nlm.nih.gov/Structure>

NCBI Conserved Domain Search

www.ncbi.nlm.nih.gov/Structure/odd/wrpsb.cgi

NCBI

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioSystems

### Conserved Domains

Search for Conserved Domains within a protein or coding nucleotide sequence

**NEW!** Use Batch CD-search to submit multiple query proteins at once!

Enter **protein** or **nucleotide** query as accession, gi, or sequence in **FASTA format** ?

```
>NP_005206.1 deleted in colorectal carcinoma [Homo sapiens]
MENSLRCVWPKLAFVLFQASLLSAHLQVTGFOIKAPTALRFLSEPSDAVTMRGGNVLLDCSAESDRGVP
VIKWKKGDIHLALGMDERKQQLSNGSLLIQNLHSRHHKPDGLYQCEASLGDSGSIISRTAKVAVAGPL
RFLSQTESVTAFMGDTVLLKCEVIGEPMPMTIHQKNNQDLTPIPGDSRVVVLPSGALQISRLQPGDIGIY
RCSARNPASSRTGNEAEVRILSDPGLHRQLYFLQRPNSVVAIEGKDAVLECCVSGYPPPSFTWLRGEEVI
QLRSKKYSLLGGSNLLISNVTDDSGMYTCVVITYKNENISASAEITVLVPPWFLNHPNLYAYESMDIEF
ECTVSGKPVPTVNMKNGDVVIPSDYFQIVGGSNLRILGVVKSDEGFYQCAEENAGNAQTSACLIVPKP
AIPSSSVLPSAPRDVVPVLVSSRFVRLSWRPPAEAKGNIQTFTVFFSREGDNRRALNTTQPGSLQLTVG
NLKPEAMYTFRVVAYNEWGPGESSQPIKQVATQPELQVPGPVENLQAVSTSPSTILITWEPAYANGPVQG
YRLPCTEVSTGKEQNIIEVDGLSYKLEGLKFTYSRLFLAYNRYGPGVSTDDITVVTLSDVPSAPPQNV
LEVNSRSIKVSWLPPSGTQNGFITGYKIRHRTTRRGEMETLEPNNLWYLFTELEKGSQYSFQVSAMT
VNGTGPPSNWYTAETPENDLDESQVPDQPSLHVRPQTNCIIMSWTPPLNPNIVVRGYIIGYGVGSPYAE
TVRVDSKQRYYSIERLESSSHYVISLKAFFNAGEGVPLYESATTSITDPTDPVDYYPLLDDFPTSVPDL
STPMLPPVGQVAVALTHDAVRVSWADNSVPKNQKTSEVRLYTVRWRTSFSASAKYKSEDTTSLSYTATGL
KPNTMYEFSVMVTKNRRSSTWSMTAHATTYEAAPTSA PKDFTVITREGKPRIVISWQPPLEANGKITAY
ILFYTLDKNIPIDDWIMETISGDRLTHQIMDLNLDITMYFRIQARNKSGVGLSDPILFRTLKVEHPDKM
```

**Submit** **Reset**

**OPTIONS**

Search against database ? : CDD v3.14 - 47363 PSSMs

Expect Value ? threshold: 0.010000

Apply low-complexity filter ? ☐

Composition based statistics adjustment ? ☒


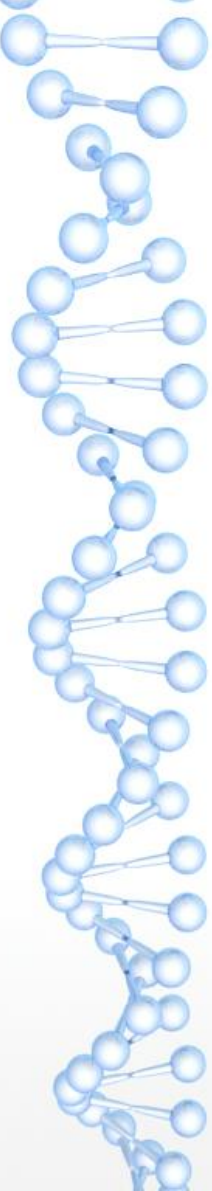
Force live search ? ☐

Rescue borderline hits ☐ Suppress weak overlapping hits ☐

Maximum number of hits ? 500

Result mode ☒ Concise ? ☐ Standard ? ☐ Full ?


# InterPro <https://www.ebi.ac.uk/interpro/>



## InterPro

Classification of protein families

Home Search Browse Results Release notes Download Help About



### Classification of protein families

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

**InterPro 76.0**  
16 September 2019

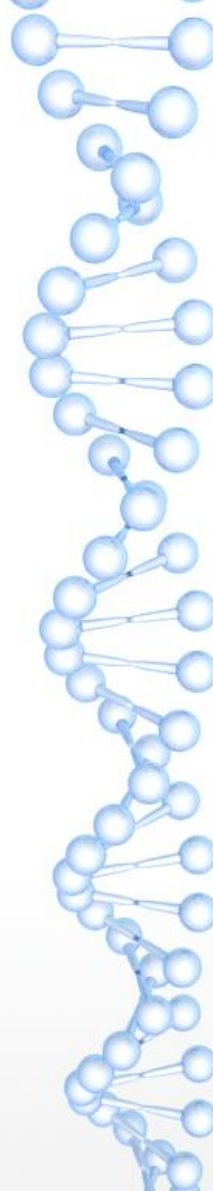
Search by sequence Search by text Search by Domain Architecture

Sequence, in FASTA format

Enter your sequence

Choose file Example protein sequence





# psi-blast

- Position-Specific Iterated BLAST search
- Used to identify distantly related sequences that are possibly missed during a standard BLAST search
- Easy-to-use version of a profile-based search
  - Perform BLAST search against protein database
  - Use results to calculate a position-specific scoring matrix
  - PSSM replaces query for next round of searches
  - May be iterated until no new significant alignments are found

**BLAST®** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastp suite

### Standard Protein BLAST

blastn blastp blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

#### Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

MKGDDPKKPKGKMSYAFVQTCREHKKKHPDASVNFSEFSKKCSERKWTNSAKERKGFEDMAKADKAR  
YEREMKTYIPPGKSTKKKPKDPNAPKPPSAFLPCEYRFPKIKGSHPLSLGDPVAKKLGEMNNTAAD  
KQPYEKKAARKLKEKYKGDIAAYRAKGRPDAAKGVYKAKSKKKKEEDEDDEDEDEDEDEDEDEDE  
DDDDDE

[Clear](#) [Query subrange](#) [?](#)

From   
To

Or, upload file  No file selected. [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

#### Choose Search Set

Database  [?](#) [?](#)

Organism  ☐ Exclude [+](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude ☒ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query

Enter an Entrez query to limit search [?](#) [YouTube](#) [Create custom database](#)

#### Program Selection


Algorithm

- ☐ blastp (protein-protein BLAST)
- ☒ PSI-BLAST (Position-Specific Iterated BLAST)
- ☐ PHI-BLAST (Pattern Hit Initiated BLAST)
- ☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

**BLAST** Search database Reference proteins (refseq\_protein) using PSI-BLAST (Position-Specific Iterated BLAST)

☐ Show results in a new window



**Algorithm parameters** [?](#) **Note:** Parameter values that differ from the default are highlighted in yellow and marked with [?](#)

#### General Parameters

Max target sequences  [?](#)  
Select the maximum number of aligned sequences to display [?](#)  
Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).

Short queries ☒ Automatically adjust parameters for short input sequences [?](#)

**Expect threshold**  [?](#) **Default = 10**

Word size  [?](#) [?](#)

Max matches in a query range  [?](#)

#### Scoring Parameters

Matrix  [?](#) [?](#)

Gap Costs  [?](#) [?](#)

Compositional adjustments  [?](#) [?](#)

#### Filters and Masking

Filter ☒ Low complexity regions [?](#)

Mask ☐ Mask for lookup table only [?](#)  
☐ Mask lower case letters [?](#)

#### PSI/PHI/DELTA BLAST

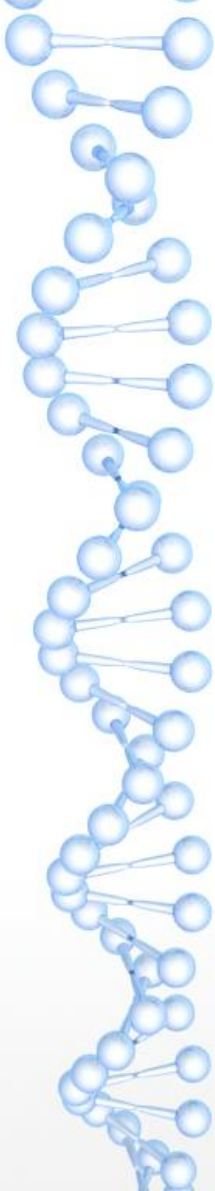
Upload PSSM  No file selected. [?](#)

**PSI-BLAST Threshold**  [?](#) **Default = 0.005**

Pseudocount  [?](#)

**BLAST** Search database Reference proteins (refseq\_protein) using PSI-BLAST (Position-Specific Iterated BLAST)

☐ Show results in a new window



NCBI BlastNP\_002119.1 hi... x

blast.ncbi.nlm.nih.gov/Blast.cgi

Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - D3Z9RUGM015

Your search is limited to records matching entrez query: (txid33511 [ORGN]) NOT((XP\_000001:XP\_999999[pacc] OR XP\_000000001:XP\_99999999[pacc])).

Edit and Resubmit Save Search Strategies > Formatting options > Download

PSI blast Iteration 9

NP\_002119.1 high-mobility group box 1 [Homo]

RID D3Z9RUGM015 (Expires on 02-29 02:47 am)

Query ID lcl|Query\_52325

Description NP\_002119.1 high-mobility group box 1 [Homo sapiens]

Molecule type amino acid

Query Length 215

Database Name refseq\_protein

Description NCBI Protein Reference Sequences

Program BLASTP 2.3.1+ > Citation

No new sequences were found above the 0.001 threshold

Other reports: > Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

Graphic Summary

Distribution of 156 Blast Hits on the Query Sequence

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

Color key for alignment scores

Score Range	Color
<40	Black
40-50	Blue
50-80	Green
80-200	Pink
>=200	Red

Query 1 40 80 120 160 200

117

156

Check Statistics



## DELTA-BLAST

- Method different from that used by PSI-BLAST

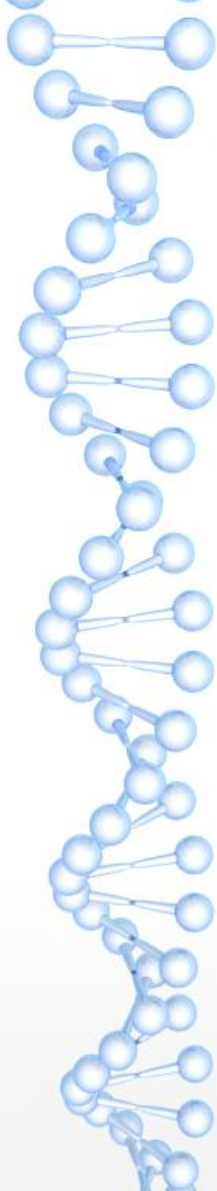
Step 1: Align the query against conserved domains derived from CDD

Step 2: Compute PSSM

Step 3: Search sequence databases using PSSM as the query

- Intended to improve homology detection
- Produces high-quality alignments, even at low levels of sequence similarity
- Dependent on homologous relationships captured within CDD





# Multiple sequence alignment: a quick primer

## Why do multiple sequence alignments?

- Identify conserved regions, patterns, and domains
  - Experimental design
  - Predicting structure and function
  - Identifying new members of protein families
- Provide basis for:
  - Predicting secondary structure
  - Performing phylogenetic analyses, thereby determining evolutionary relationships (inferring homology)
  - Generating position-specific scoring matrices for use with sensitive sequence search methods





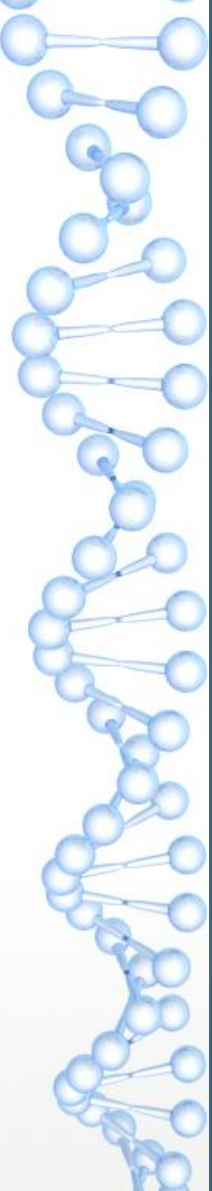
## Overarching Considerations

- Absolute sequence similarity  
*Create the alignment by lining up as many common characters as possible*
- Conservation  
*Take into account residues that can substitute for one another and not adversely affect the function of the protein*
- Structural similarity  
*Knowledge of the secondary or tertiary structure of the proteins being aligned can be used to fine-tune the alignment*



## Protein vs. Nucleotide Multiple Sequence Alignments

- Concentrate on the protein level rather than on the nucleotide level
- Protein alignments tend to be more informative
- Less prone to inaccurate alignment ('20 vs. 4')
- Can 'translate back' to nucleotide sequences *after* doing the alignment



*Find sequences to align through database searches  
satisfying a reasonable E-value cutoff*



*Run the multiple sequence alignment program*



*Inspect and assess the quality of the alignment*



*Remove sequences that seriously disrupt the alignment, then realign*



*Add back remaining sequences, based on key residues in the alignment*



*Interpret the alignment*





# Selecting the Sequences

1. Use a reasonable number of sequences to avoid technical difficulties
  - **Global** alignment method: compute time increases exponentially as sequences are added to the set
  - Most alignment algorithms are ineffective on huge data sets (and may yield inaccurate alignments)
  - Phylogenetic studies resulting from inordinately large data sets can sometimes be intractable
  - Good starting point: 10-15 sequences
  - Ballpark upper limit: 50-100 sequences



# Selecting the Sequences

2. Sequences should be of about the same length
3. Trim sequences down, so as to only use regions that have been deemed similar by either:
  - Pairwise search methods such as BLAST
  - Profile-based search methods such as PSI-BLAST





# Selecting the Sequences

4. Consider the degree of similarity in the sequence set,  
*depending on what question is being asked*
  - Use closely-related sequences to determine 'required' (highly conserved) amino acids
  - Use more divergent sequences to study evolutionary relationships
  - Good starting point: use sequences that are 30-70% similar to most of the other sequences in the data set
  - The most informative alignments result when the sequences in the data set are not too similar, but also not too dissimilar



# Inspection: an iterative process

- Perform alignment on small set of sequences
- Examine the quality of the alignment, looking for:
  - Conservation of residues across alignment
  - Conservation of physicochemical properties
  - Relatively neat block-type structure
  - Excessive numbers of gaps
- If alignment is good, can add new sequences to data set, then realign
- If alignment is not good, remove any sequences that result in the inclusion of long gaps, then realign



# Inspection: an iterative process

- Use visualization tools to identify 'key residues' and 'problem regions'
- Cross-check against 'expertly created' multiple sequence alignments available online
- Use any available information from solved X-ray or NMR structures to nail down structurally important regions and to assess where gaps can (or cannot) be tolerated





# Interpretation

## Inspection: An Iterative Process

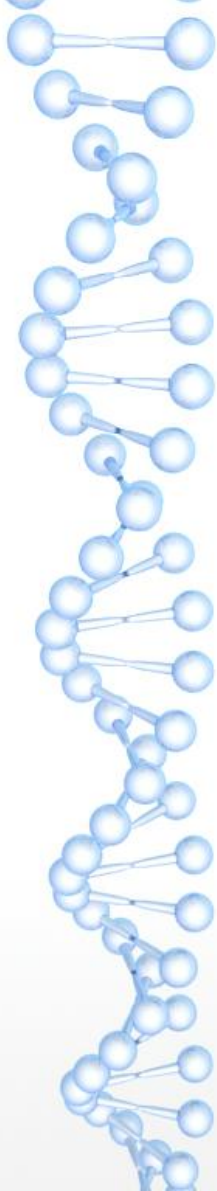
- Use visualization tools to identify 'key residues' and 'problem regions'
- Cross-check against 'expertly created' multiple sequence alignments available online
- Use any available information from solved X-ray or NMR structures to nail down structurally important regions and to assess where gaps can (or cannot) be tolerated



# Clustal omega

- Allows for automatic multiple alignment of nucleotide or amino acid sequences
- Aligns data sets quickly and easily
- Can align sequences against a pre-existing alignment (an 'external profile')
- Can bias the location of gaps, based on known structural information
- Works with Jalview, a Java applet for viewing and manipulating results





- Align two sequences at a time, starting with the two most related sequences
- Gradually build up the multiple sequence alignment by adding additional (less-related) sequences to the alignment
- Uses protein scoring matrices and gap penalties to calculate alignments having the best score
- Major advantages of method
  - Generally fast
  - Alignments generally of high quality



# Clustal omega output

- Pairwise alignment scores
- Multiple sequence alignment
- Cladogram
  - Tree that is assumed to be an *estimate* of a phylogeny
  - Branches are of equal length
  - Cladograms can show common ancestry, but do not provide an indication of the amount of evolutionary time separating taxa
- Phylogram
  - Tree that is assumed to be an *estimate* of a phylogeny
  - Branches are *not* of equal length
  - Branch lengths proportional to the amount of inferred evolutionary change



# Conservation pattern

*Conservation patterns in multiple sequence alignments usually follow the following rules:*

[ WYF ]	Aromatics
[ KRH ]	Basic side chains (+)
[ DE ]	Acidic side chains (–)
[ GP ]	Ends of helices
[ HS ]	Catalytic sites
[ C ]	Cysteine cross-bridges



# Conservation pattern

*Interpretation is empirical — there is no parallel to the E-values seen in BLAST searches to assess statistical significance*

- \* entirely conserved column  
(want in at least 10% of positions)
- ⋮ conserved  
(strongly similar properties)
- semi-conserved  
(weakly similar properties)



# Clustal Omega

Input form Web services Help & Documentation

Share Feedback

Tools > Multiple Sequence Alignment > Clustal Omega

## Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

### STEP 1 - Enter your input sequences

Enter or paste a set of **PROTEIN** sequences in any supported format:

```
>FOSB_MOUSE Protein fosB
MFQAFPGDYDSGSRCS SSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLQPTLISSMAQSQQGPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPV SARPARARPRRPREETLTPEEEEEKRRVRRERNKLAAKCRNRRREL
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTA SLFTHSEVQVLGDPFPVVS PSY
TSSFVLTCPEVSFAFAGAQTSGSEQPSDPLNSPLLAL
```

Or, upload a file: **Browse...** No file selected.

### STEP 2 - Set your parameters

OUTPUT FORMAT **Clustal w/o numbers**

The default settings will fulfill the needs of most users and, for that reason, are not visible.

**More options...** (Click here, if you want to view or change the default settings.)

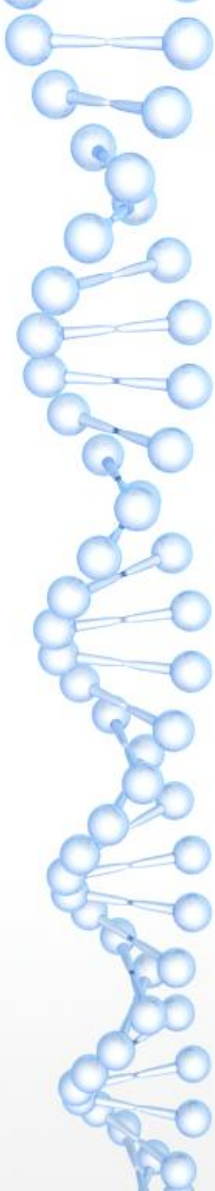
### STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

**Submit**

If you plan to use these services during a course please [contact us](#).

Please read the FAQ before seeking help from our support staff.



STEP 2 - Set your parameters

OUTPUT FORMAT Clustal w/ numbers

DEALIGN INPUT SEQUENCES	MBED-LIKE CLUSTERING GUIDE-TREE	MBED-LIKE CLUSTERING ITERATION	NUM
yes	yes	yes	defa
MAX GUIDE TREE ITERATIONS	MAX HMM ITERATIONS	ORDER	
default	default	input	

STEP 3 - Submit your job

☐ notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

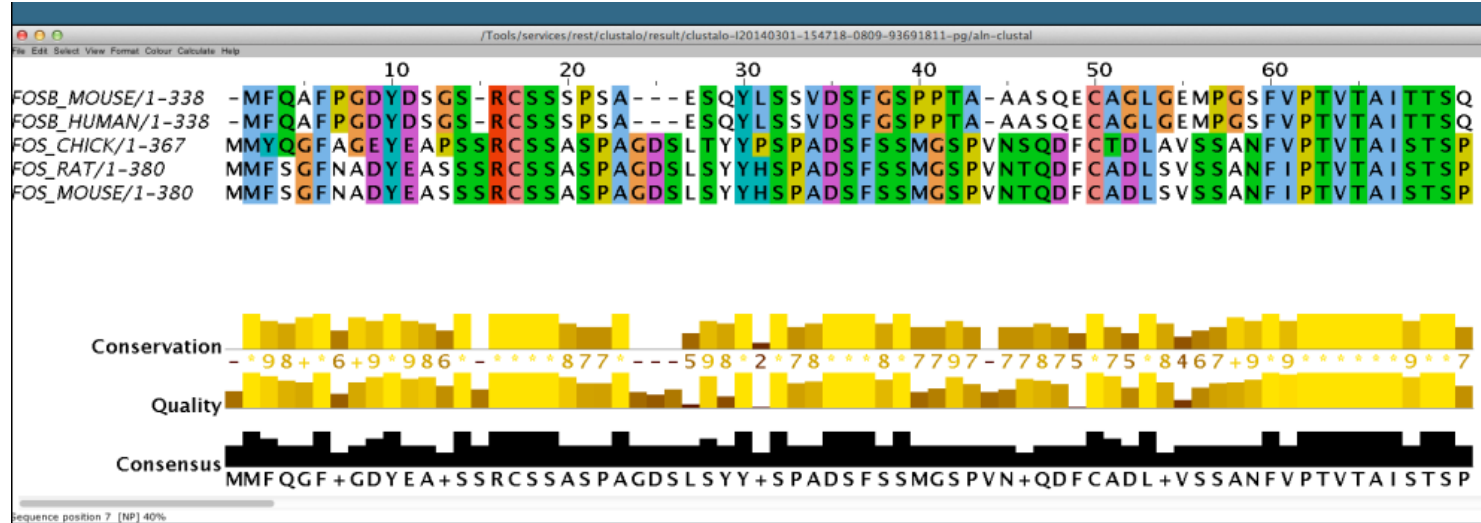
If you plan to use these services during a course please [contact us](#).



# Jalview

- Java applet available within Clustal Omega results
- Used to manually edit Clustal Omega alignments
- Color residues based on various properties
- Pairwise alignment of selected sequences
- Consensus sequence calculations
- Removal of redundant sequences
- Calculation of phylogenetic trees

# Jailview



## Conservation

Conservation of total alignment  
(indication of percent identity)

## Quality

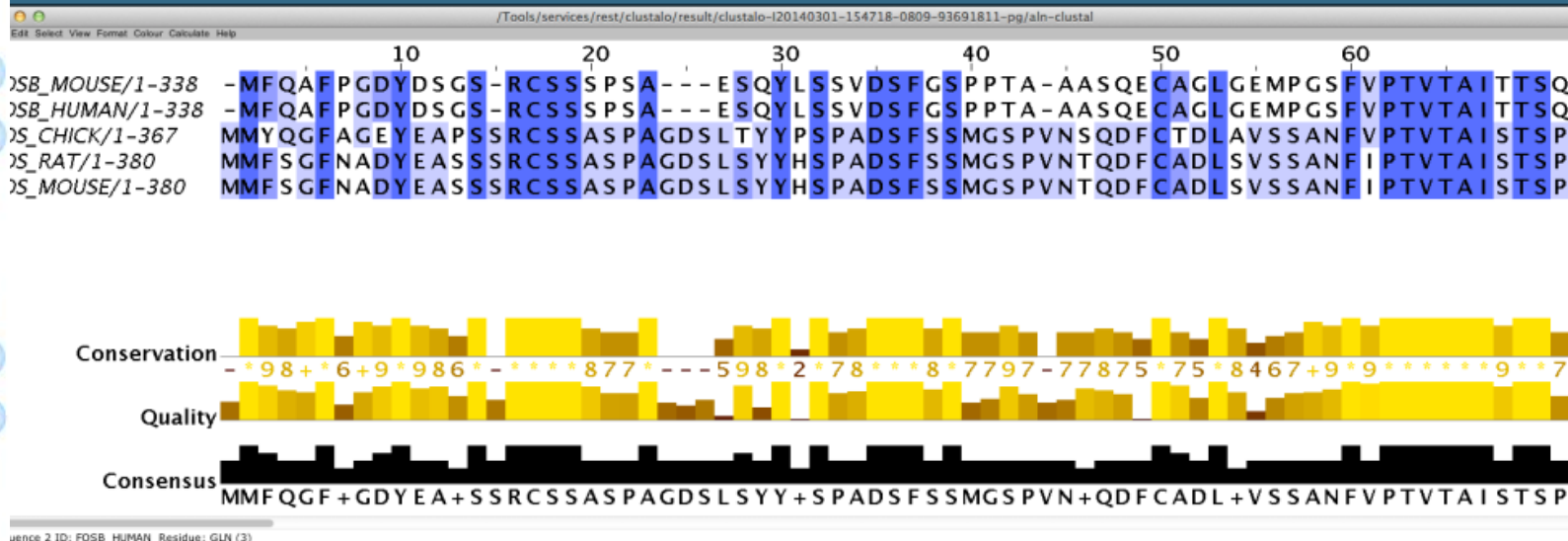
Alignment quality, based on BLOSUM scores

## Consensus

Based on percent identity



## Colour → Percentage Identity



### Agreement

### Background Color

81 - 100%

Dark blue

61 - 80%

Medium blue

41 - 60%

Light blue

≤ 40%

White

## T-COFFEE

- Combines sequence, profile, and structural information
  - Protein structures
  - RNA secondary structures
- Specialized algorithm for aligning transmembrane proteins, non-coding RNAs, and homologous promoter regions
- Can combine output from other methods into a single 'master alignment'
- Freely available at <http://tcoffee.org>



Magis et al., *Methods Mol. Biol.* 1079: 117-129 (2014)

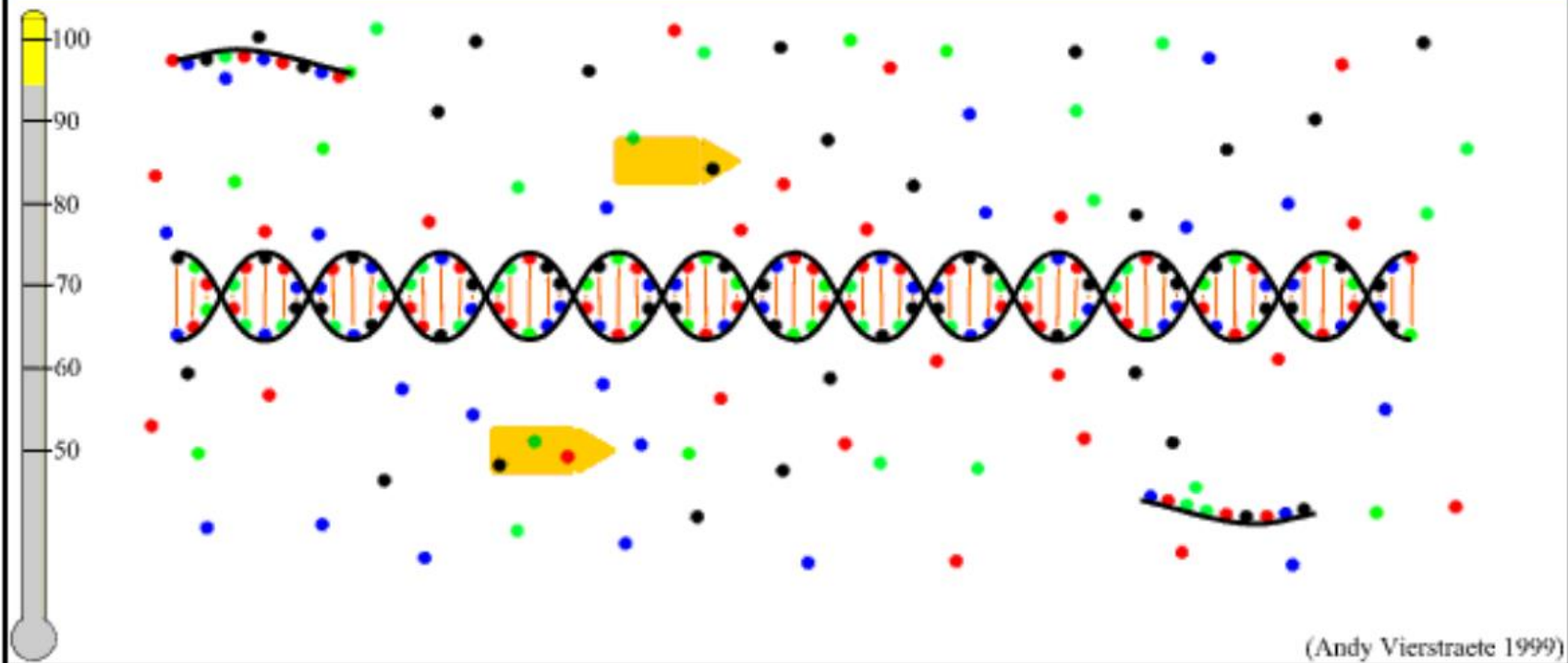


A **primer** is a strand of short nucleic acid sequences that serves as a starting point for **DNA synthesis**. It is required for DNA replication because the enzymes that catalyze this process, **DNA polymerases**, can only add new nucleotides to an existing strand of DNA. The polymerase starts replication at the 3'-end of the primer, and copies the opposite strand.

### Reverse primer

PCR :

Denaturation 94°C







# Primer design

For a successful and reliable PCR requires efficient and specific amplification of the product, using chemically synthesized oligonucleotides – DNA primers.

Target sequence and designing primers substantially affect the efficiency of your PCR

When designing primers, follow these steps:

1. Check literature and databases for existing primers
2. Choose a target sequence
3. Design primers (and probes)
4. Check primer specificity
5. Validate primers



# Primer design

## Target Sequence for PCR

Plan to amplify

Conventional PCR: **200-800 bp** product (~500)

Real Time PCR: **75-200 bp** (~100)

Short PCR products are typically amplified with higher efficiency than longer ones; but should be at least 75 bp to easily distinguish from any primer-dimers



# Primer design

## Target Sequence for PCR

Plan to amplify

Conventional PCR: **200-800 bp** product (~500)

Real Time PCR: **75-200 bp** (~100)

Short PCR products are typically amplified with higher efficiency than longer ones; but should be at least 75 bp to easily distinguish from any primer-dimers



# Primer design

## Uniqueness

There shall be one and only one target site in the template DNA where the primer binds, which means the **primer sequence shall be unique in the template DNA**, avoiding the possibility of mishybridization to a similar sequence nearby.

There shall be no annealing site in possible **contaminant sources**, such as human, rat, mouse, etc. (BLAST search against corresponding genome)

**Verify specificity** using tools such as the Basic Local Alignment SearchTool (BLAST) (<http://www.ncbi.nlm.nih.gov/blast/>)





# Primer design

## Length

Primer length has effects on **uniqueness and melting/annealing temperature**. Roughly speaking:

- the longer the primer, the more chance that it is **unique**;
- the longer the primer, the higher melting/annealing temperature – **specificity**

The length of primer has to be at least 15 bases to ensure uniqueness. Usually, we pick primers of **17-28** bases long.

This range varies based on if you can find unique primers with appropriate annealing temperature within this range.

Above 30: risk of mispairing, primer dimers, and hairpins



# Primer design

**Base composition** affects hybridization specificity and melting/annealing temperature.

- **Random** base composition is preferred!

Avoid long (A+T) and (G+C) rich region if possible. Repeats: A repeat is a di-nucleotide occurring many times consecutively and should be avoided because they can misprime. A maximum number of di-nucleotide repeats acceptable in an oligo is 4 di-nucleotides.

Template DNA 5' ..TCG**ATATATAT**GCATG...GAT**GCCGGCGCGC**TGTACACAA..3'

Primers with long runs of a single base should generally be avoided as they can misprime. For example, AGC**GGGGG**AT**GGGG** has runs of base 'G' of value 5 and 4. Avoid repeats of more than 3 bases - 4 bases is accepted

- Usually, **average (G+C) content around 40-60%** will give us the right melting/annealing temperature for ordinary PCR reactions, and will give appropriate hybridization stability.



# Primer design

## Melting Temperature - $T_m$

The temperature at which 50% of the primer molecules are bound to their corresponding target sequence.

$T_m$  is characteristics of the DNA/Base composition; Higher G+C content DNA, has a higher  $T_m$  due to more **Hydrogen-bonds**.

3 vs. only 2 in A::T

*Calculation*

$$T_m = 64.9 + 41 * (yG + zC - 16.4) / (wA + xT + yG + zC)$$

(Formulae are from <http://www.basic.northwestern.edu/biotools/oligocalc.html>)



# Primer design

## Annealing Temperature

Annealing Temperature,  $T_{\text{anneal}}$  – the temperature at which primers anneal to the template DNA. It can be calculated from  $T_m$ .

$$T_{\text{anneal}} = T_{m\_primer} - 4^{\circ}\text{C}$$

**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

Mismatch tolerance is found to have the strongest influence on PCR specificity

The optimal  $T$  for PCR often needs to be determined empirically



# Primer design

If primers can anneal to themselves or anneal to each other (**primer dimer**) rather than anneal to the template, the PCR efficiency will be decreased dramatically. They shall be avoided.

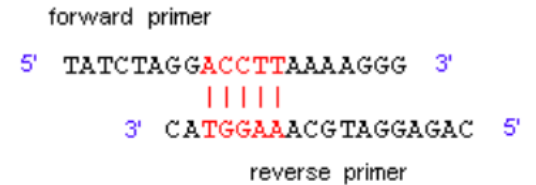
## Hairpin



## Self-Dimer



## Dimer



However, sometimes these 2° structures are harmless when the annealing temperature does not allow them to take form. For example, some dimers or hairpins form at 30°C while during PCR cycle, the lowest temperature only drops to 60°C.



# Primer design

Primers work in pairs – **forward** primer and **reverse** primer. Since they are used in the same PCR reaction, you should make sure that the *PCR condition is suitable for both of them*.

One critical feature is their annealing temperatures, which shall be compatible with each other.

The maximum **difference allowed is 3°C**. The closer their  $T_{\text{anneal}}$  are, the better.

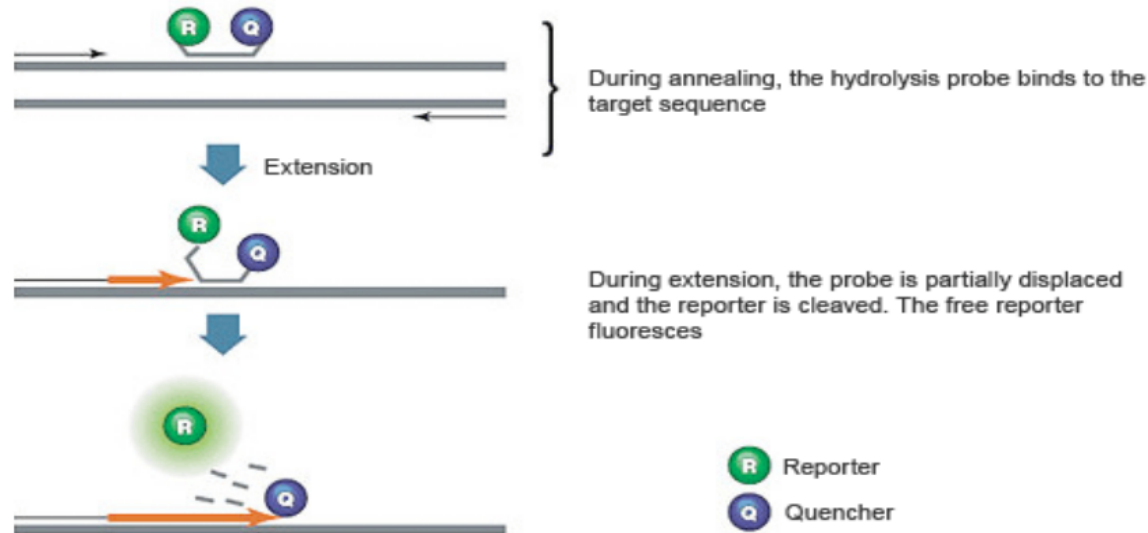
# Primer design

## Probes for Real Time PCR – greater specificity

The  $T_m$  of the probe should be 5–10°C higher than that of the primers

In most cases, the probe should be <30 nucleotides

Choose a sequence within the target that has a GC content of 30–80% - uniqueness and specificity still apply – minimise the risk of mispriming





# Primer design: summary

1. Uniqueness: ensure correct priming site
2. Length: 17-28 bases. This range varies
3. Base composition: average (G+C) content around 40-60%; avoid long (A+T) and (G+C) rich region if possible
4. Optimize base pairing: G or C in the 3' end but not too many - to minimize false priming
5. Melting  $T_m$  between 52-65°C are preferred
6. Assure that F/R primers have annealing  $T$  within 2 – 3 °C of each other
7. Minimize internal secondary structure: hairpins and dimers shall be avoided (minimize self complementarity and 3' end self complementarity)





# Primer design

## Multiplex PCR

Multiple primer pairs can be added in the same tube amplify **multiple sites**

Application example: genome identification

## Design difficulty

- Similar melting Temperature
- No dimer formulation (cross-dimer)
- The products need to be of different sizes if visualization by gel – or use different probes/fluophores



# Primer design

Primers can also be designed to amplify **multiple products** - “universal primers”.

For example, design primers to amplify all Dengue serotypes.

Strategy:

1. Align groups of sequences you want to amplify.
2. Find the most conservative regions at 5' end and at 3' end.
3. Design forward and reverse primers and find the best matching pair.
4. Ensure uniqueness in all template sequences.



# Primer design

**Free internet resources for designing primers and probes:**

**Primer3 (Whitehead Institute, MIT)**

<http://bioinfo.ut.ee/primer3/>

GeneFisher (Bielefeld University)

<http://bibiserv.techfak.uni-bielefeld.de/genefisher/>

FastPCR (Biocenter, University of Helsinki)

<http://www.biocenter.helsinki.fi/bi/Programs/fastpcr.htm>

PerlPrimer (Owen Marshall)

<http://perlprimer.sourceforge.net/>

Primer Design Assistant (Division of Biostatistics and Bioinformatics, NHRI)

<http://dbb.nhri.org.tw/primer/>

Melting temperature calculation software:

- **BioMath**
- **Applied Biosystems**

# Primer design

<b>Primer3</b> (v. 0.4.0) Pick primers from a DNA sequence.	<a href="#">Checks for mispriming in template.</a>	<a href="#">disclaimer</a>	<a href="#">Primer3 Home</a>
	<a href="#">Primer3plus interface</a>	<a href="#">cautions</a>	<a href="#">FAQ/WIKI</a>

Paste source sequence below (5'→3', string of ACGTNacgtn -- other letters treated as N -- numbers and blanks ignored). FASTA format ok. Please N-out undesirable sequence (vector, ALUs, LINEs, etc.) or use a [Mispriming Library \(repeat library\)](#):

```
>PF13_0006
ATGAAACTTCACCTGCTCTAAATATTATTATTTTACTTCCATTAAATATATTAGTAACA
TCATTATCAAATGTGCATAGTAAAAATAAACCATACATCACATCACGTCATACAGCAACT
ACTATATCGCGAGTGTTAAGCGAATGTGACATCCGATCGTCAATTTATGATAATGATGAG
GATATCAAATCAGTGAAGGAATGTTTTGATCGACAAACATCACAACGATTTGAAGAATAC
GAAGAACGTATTCAAGAAAAACGCCAAAAACGTAAAGAAGAACGGGACAAAAATATAAAA
```

<input checked="" type="checkbox"/> Pick left primer, or use left primer below:	<input type="checkbox"/> Pick hybridization probe (internal oligo), or use oligo below:	<input checked="" type="checkbox"/> Pick right primer, or use right primer below (5' to 3' on opposite strand):
<input type="text"/>	<input type="text"/>	<input type="text"/>

[Sequence Id:](#)  A string to identify your output.

[Targets:](#)  E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the [source sequence](#) with [ and ]: e.g. ...ATCT[CCCC]TCAT.. means that primers must flank the central CCCC.

[Excluded Regions:](#)  E.g. 401,7 68,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 68. Or mark the [source sequence](#) with < and >: e.g. ...ATCT<CCCC>TCAT.. forbids primers in the central CCCC.

[Product Size Ranges:](#)

<a href="#">Number To Return</a>	<input type="text" value="5"/>	<a href="#">Max 3' Stability</a>	<input type="text" value="9.0"/>
<a href="#">Max Repeat Mispriming</a>	<input type="text" value="12.00"/>	<a href="#">Pair Max Repeat Mispriming</a>	<input type="text" value="24.00"/>
<a href="#">Max Template Mispriming</a>	<input type="text" value="12.00"/>	<a href="#">Pair Max Template Mispriming</a>	<input type="text" value="24.00"/>



# Primer design

## General Primer Picking Conditions

<u>Primer Size</u>	Min: 18	Opt: 20	Max: 27	
<u>Primer Tm</u>	Min: 57.0	Opt: 60.0	Max: 63.0	<u>Max Tm Difference:</u> 100.0 <u>Table of thermodynamic parameters:</u> Breslauer et al. 1986 ▼
<u>Product Tm</u>	Min:	Opt:	Max:	
<u>Primer GC%</u>	Min: 40.0	Opt:	Max: 60.0	
<u>Max Self Complementarity:</u>	8.00	<u>Max 3' Self Complementarity:</u>	3.00	
<u>Max #N's:</u>	0	<u>Max Poly-X:</u>	5	
<u>Inside Target Penalty:</u>		<u>Outside Target Penalty:</u>	0	<u>Note: you can set Inside Target Penalty to allow primers inside a target.</u>
<u>First Base Index:</u>	1	<u>CG Clamp:</u>	0	
<u>Concentration of monovalent cations:</u>	50.0	<u>Salt correction formula:</u>	Schildkraut and Lifson 1965 ▼	
<u>Concentration of divalent cations</u>	0.0	<u>Concentration of dNTPs</u>	0.0	
<u>Annealing Oligo Concentration:</u>	50.0	(Not the concentration of oligos in the reaction mix but of those annealing to template.)		
<input checked="" type="checkbox"/> <u>Liberal Base</u>	<input type="checkbox"/> <u>Show Debugging Info</u>	<input checked="" type="checkbox"/> Do not treat ambiguity codes in libraries as consensus	<input type="checkbox"/> <u>Lowercase masking</u>	

Pick Primers

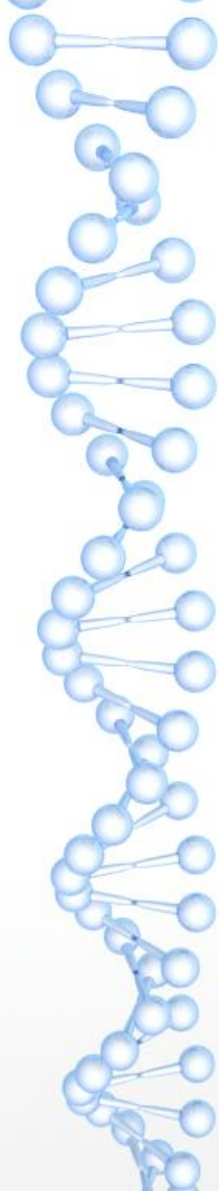
Reset Form

## 58

# Primer design

## ADDITIONAL OLIGOS

	<u>start</u>	<u>len</u>	<u>tm</u>	<u>gc%</u>	<u>any</u>	<u>3'</u>	<u>seq</u>
1 LEFT PRIMER	502	20	59.99	50.00	4.00	2.00	GCACGTATTGCGGAAGGTAT
RIGHT PRIMER	639	20	59.47	40.00	4.00	1.00	AGCACGAGCAATGTTTTTGA
PRODUCT SIZE: 138, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00							
2 LEFT PRIMER	681	20	59.93	50.00	4.00	2.00	CCCAGTCGATTCCAAACCTA
RIGHT PRIMER	812	20	59.50	50.00	5.00	1.00	TCGGCATCTGAGACGATAGA
PRODUCT SIZE: 132, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00							
3 LEFT PRIMER	253	20	60.62	40.00	2.00	2.00	CAAGAAAAACGCCAAAAACG
RIGHT PRIMER	381	20	60.03	50.00	2.00	1.00	TAACGCACACCCACACCTAA
PRODUCT SIZE: 129, PAIR ANY COMPL: 2.00, PAIR 3' COMPL: 0.00							
4 LEFT PRIMER	253	20	60.62	40.00	2.00	2.00	CAAGAAAAACGCCAAAAACG
RIGHT PRIMER	390	20	60.04	55.00	2.00	0.00	AACACCTCCTAACGCACACC
PRODUCT SIZE: 138, PAIR ANY COMPL: 2.00, PAIR 3' COMPL: 1.00							



## You still have to check for **Primer Specificity by BLAST** (NCBI)

<http://www.ncbi.nlm.nih.gov/BLAST/>

Choose a species genome to search, or list all genomic BLAST databases

[http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST\\_PROGRAMS=megaBlast&PAGE\\_TYPE=BlastSearch&SHOW\\_DEFAULTS=on&BLAST\\_SPEC=&LINK\\_LOC=blasttab&LAST\\_PAGE=blastn](http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&BLAST_SPEC=&LINK_LOC=blasttab&LAST_PAGE=blastn)


That your given primers are specific for your gene of interest only

Then you can order your primers....



Finding primers specific to your PCR template (using Primer3 and BLAST).

## PCR Template

[Reset page](#)[Save search parameters](#)[Retrieve recent results](#)[Publication](#)[Tips for finding specific primers](#)Enter accession, gi, or FASTA sequence (A refseq record is preferred) [Clear](#)

CP030219.1

Range

Forward primer

From

To

Reverse primer

[Clear](#)

Or, upload FASTA file

Sfoggia...

Nessun file selezionato.

## Primer Parameters

Use my own forward primer  
(5'→3' on plus strand)[Clear](#)Use my own reverse primer  
(5'→3' on minus strand)[Clear](#)

PCR product size

Min

70

Max

1000

# of primers to return

10

Primer melting temperatures  
(T<sub>m</sub>)

Min

57.0

Opt

60.0

Max

63.0

Max T<sub>m</sub> difference

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


7

Exon at 3' side

4

Minimal number of bases that must anneal to exons at the 5' or 3' side of the junction 

Intron inclusion

☐ Primer pair must be separated by at least one intron on the corresponding genomic DNA 

Intron length range

Min

1000

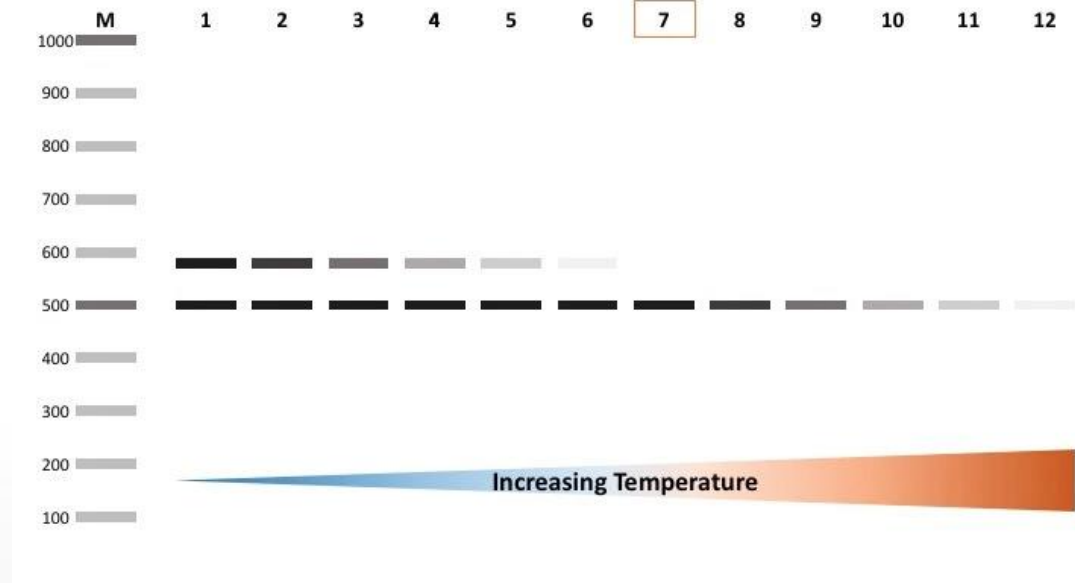
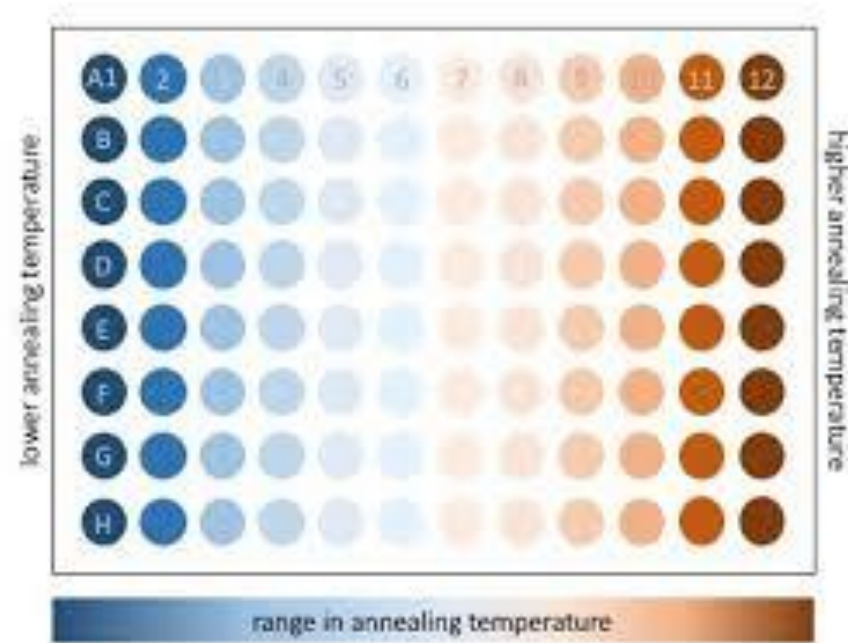
Max

1000000



# Primer design

Gradient PCR  
Optimizing pcr





# Primer design - Homework

1. Choose a pathogen

Find sequence (FASTA). Not the whole genome

NCBI - <http://www.ncbi.nlm.nih.gov/genbank/>

(PER SUGGERIMENTI: <https://www.ncbi.nlm.nih.gov/pathogens/organism>)

Copy the sequence to notepad

2. Use web-based tool to design a primer pair

Primer3 - <http://bioinfo.ut.ee/primer3/>

Paste the pathogen sequence

Fill out requirements

3. Use BLAST to test the first primer pair for off-targets

If there are off targets continue with the second primer pair

4. Design PCR program based on  $T_m$  of both primer pairs