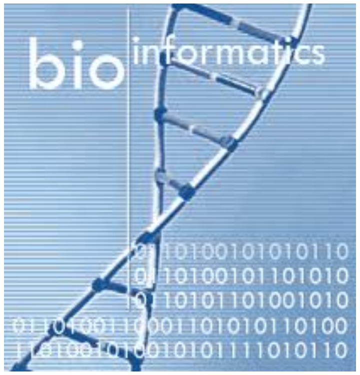
Bioinformatics

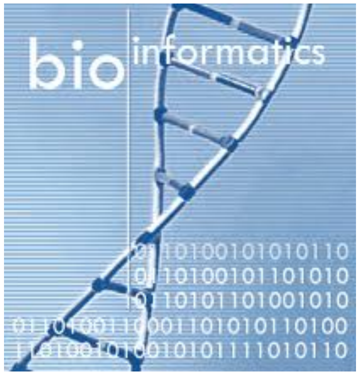
SEVEN

Next Generation Sequencing

Wendy Lee
Dept of Computer Science
San José State University



Biology/CS/SE 123A
Fall 2014



Sequencing Technologies

Traditional sequencing

- Sanger Sequencing

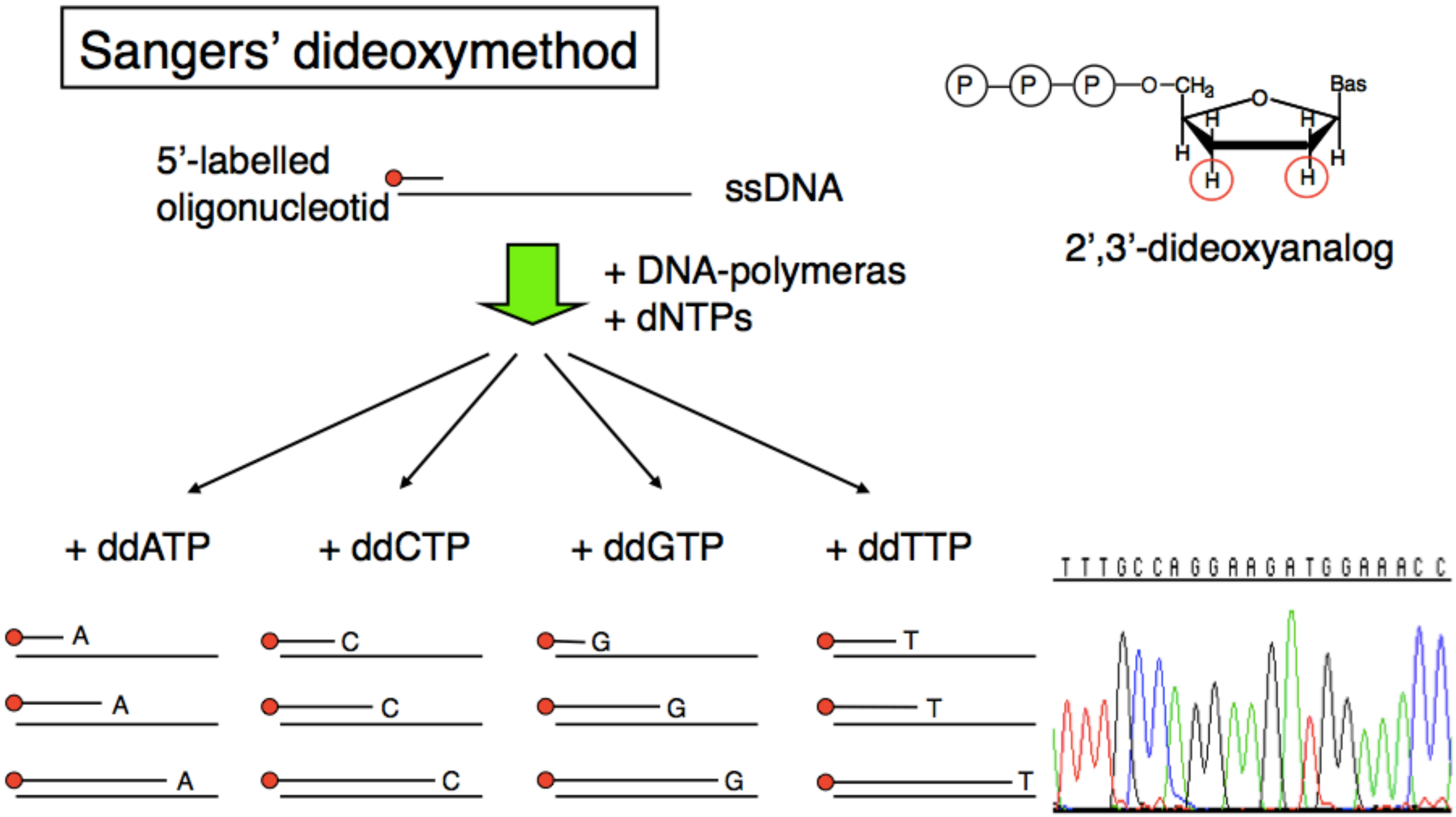
Next Generation Sequencing (NGS)

- Pyrosequencing
- Illumina/Solexa
- Ion Torrent (charge based detection)

Sanger Sequencing

- Developed by Frederick Sanger and colleagues in 1977
- It was the most widely used sequencing method for approximately 25 years.
- More recently, "Next-Generation" sequencing methods are more commonly used, especially for large-scale, automated genome analyses.
- Sanger method remains in wide use, primarily for smaller-scale projects and for obtaining especially long contiguous DNA sequence reads (>500 nucleotides).

Sanger Sequencing



Sangers' dideoxymethod

5'-labelled oligonucleotid

ssDNA

+ DNA-polymeras

+ dNTPs

+ ddATP

+ ddCTP

+ ddGTP

+ ddTTP

2',3'-dideoxyanalog

Pyrosequencing

- Pyrosequencing was developed by Mostafa Ronaghi and Pål Nyrén at the Royal Institute of Technology in Stockholm in 1996
- Pyrosequencing is a DNA sequencing technique that is based on the detection of released pyrophosphate (PPi) during DNA synthesis.

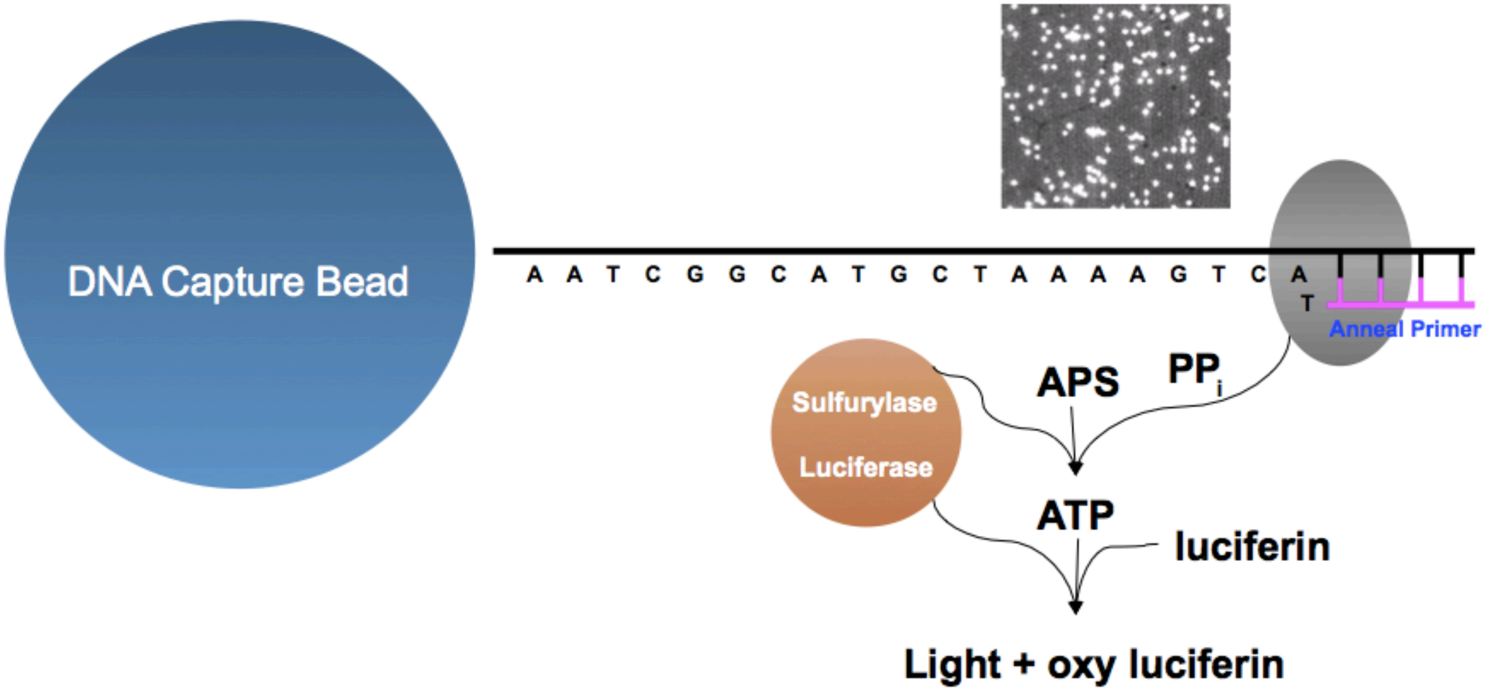
Pyrosequencing

- Sequencing-by-synthesis based
- Accurate
- Simple and robust
- No labels or gels
- Real-time results

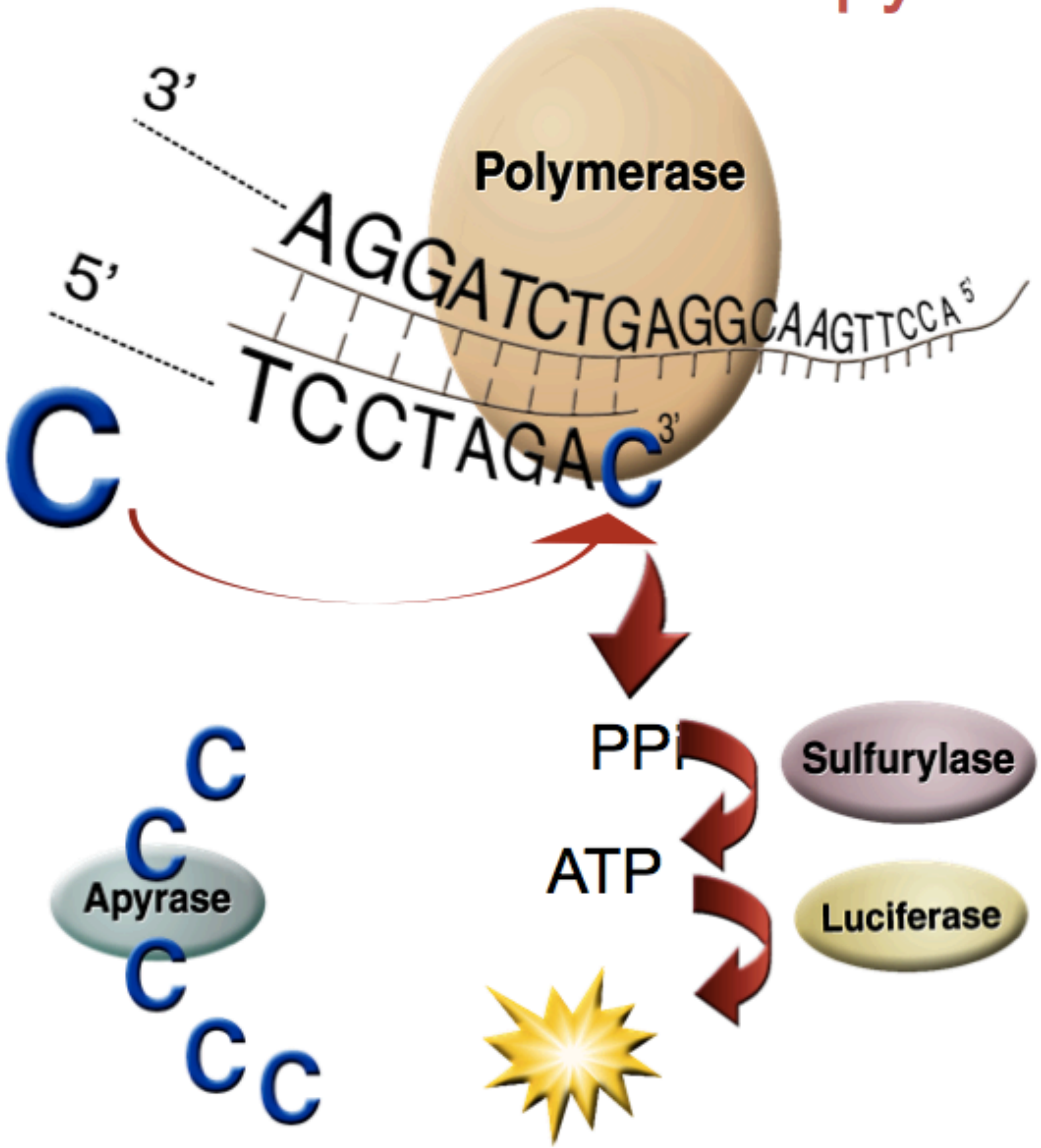
Pyrosequencing

Sequencing-By-Synthesis

- Pyrophosphate signal generation

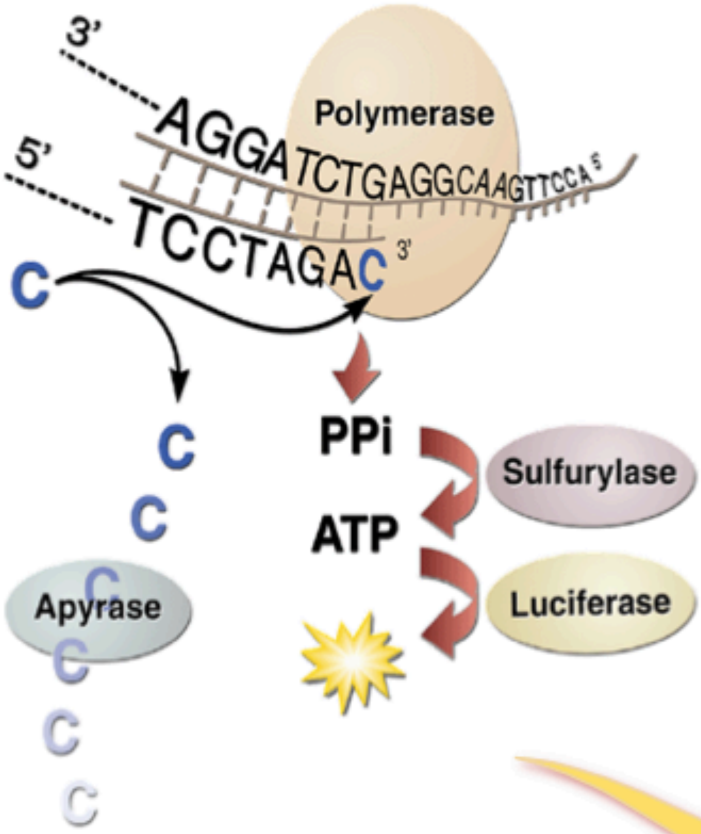


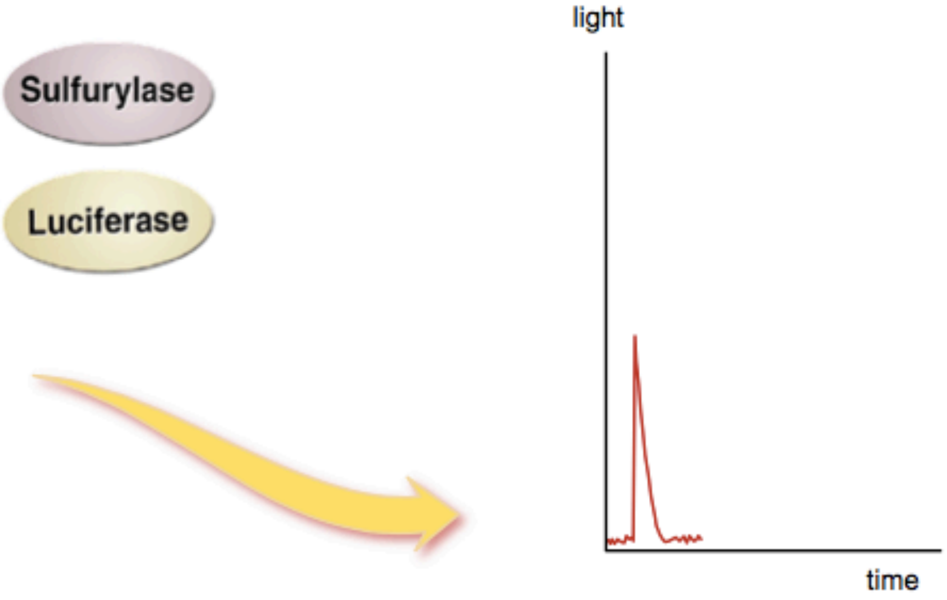
The pyrosequencing method



The pyrosequencing method

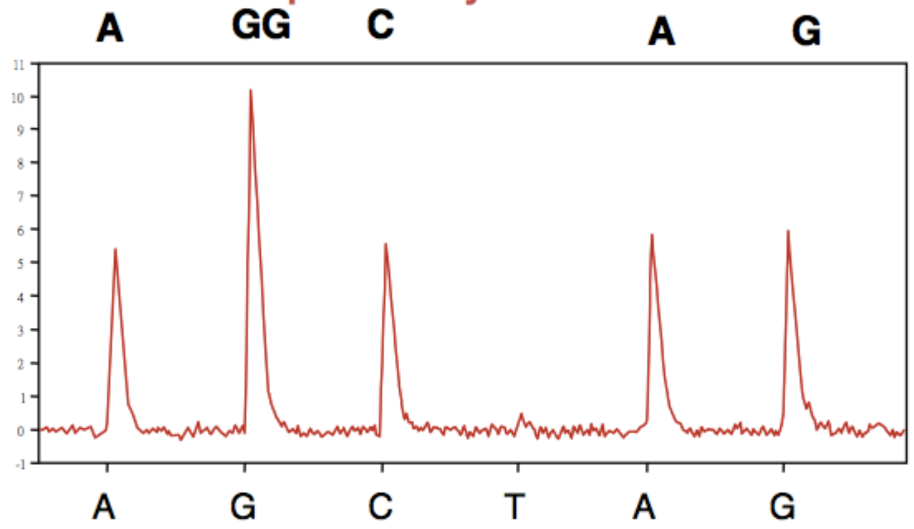
- detection of the light





The pyrosequencing method

- nucleotides dispensed sequentially



the sequence in this pyrogram™ is AGGCAG

Instrumentation

- PSQ™ 96

- Automatic dispensation of reagents
- 96 well format
- CCD camera
- Processes

  500 samples per hour

  4500 samples per day

Instrumentation

- working with the PSQ™ 96

1. Prepare samples
2. Insert samples in PSQ™ 96
3. Insert reagent cartridge (enzymes, substrate, nucleotides)
4. Start run

sequence automatically scored

Instrumentation

- PSQ **HS 96A**

- Automatic dispensation of reagents
- 96 well format (It will be possible to upgrade to a 384-format)
- CCD camera
- Processes

  10000 samples per day

  30000 samples per day

  (Triplex analysis)

Applications

- one technology, many applications

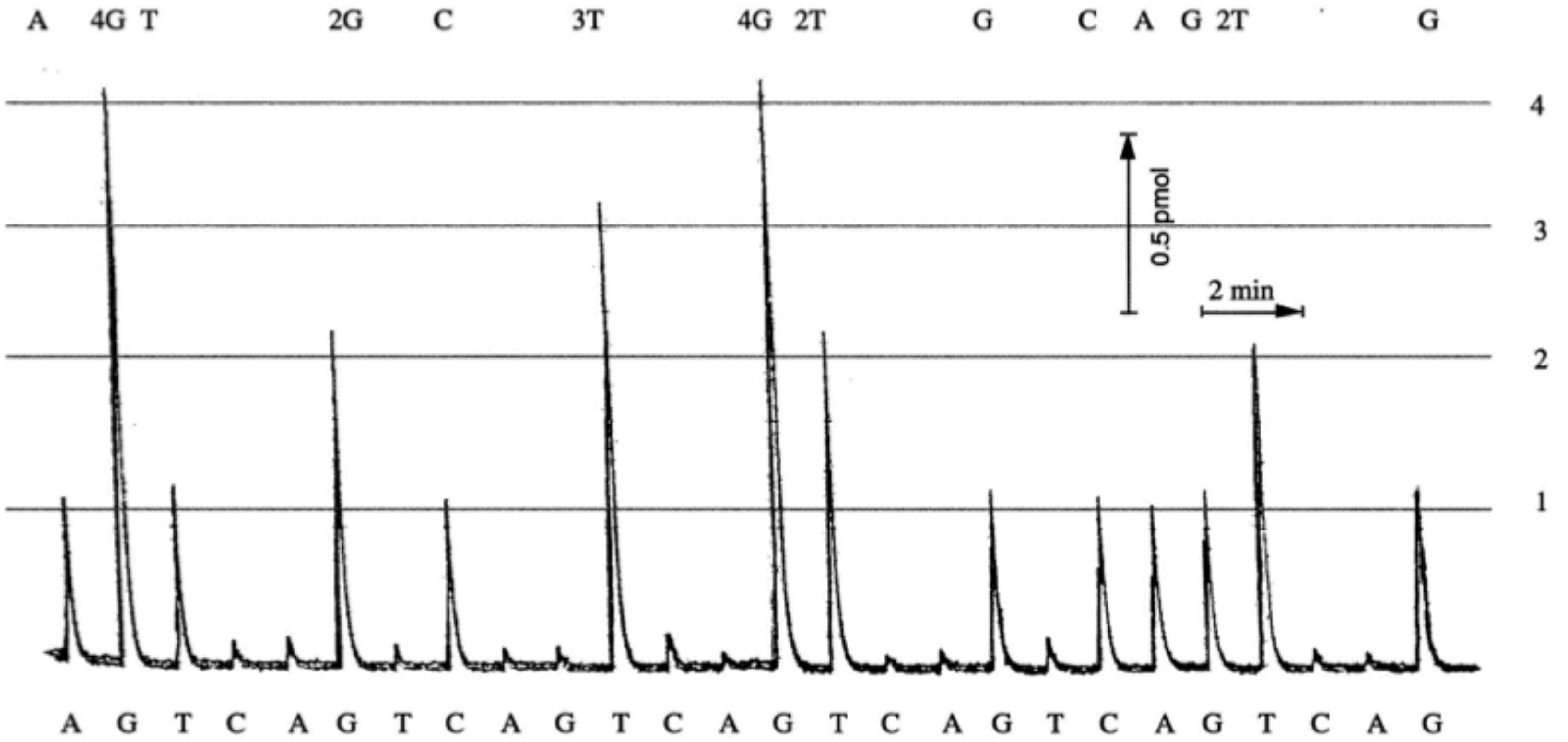
- Genetic variability (SNP, insertions, deletions)
- Haplotyping
- Allele quantification / frequency
- Expression profiling, clone identification etc.
- Bacterial and viral typing
- Resistance typing
- Mutation detection
- Forensic study ..and more

Applications - SNP Analysis

- SNPs as genetic markers

- Single Nucleotide Polymorphisms are isolated single base variations in the genome
- Occur every 500-1000 bases along the 3 billion bases of the human genome
- The most common form of genetic inter-individual variation
- The major source of phenotypic variability between individuals

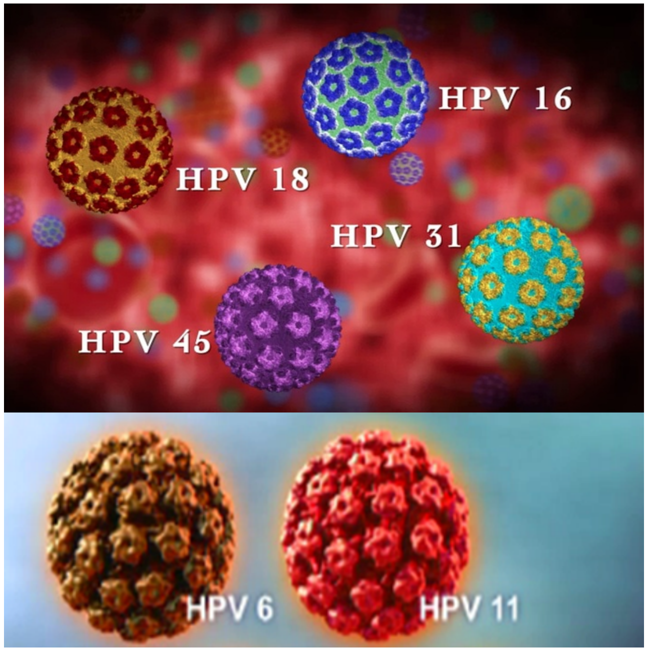
Pyrogram



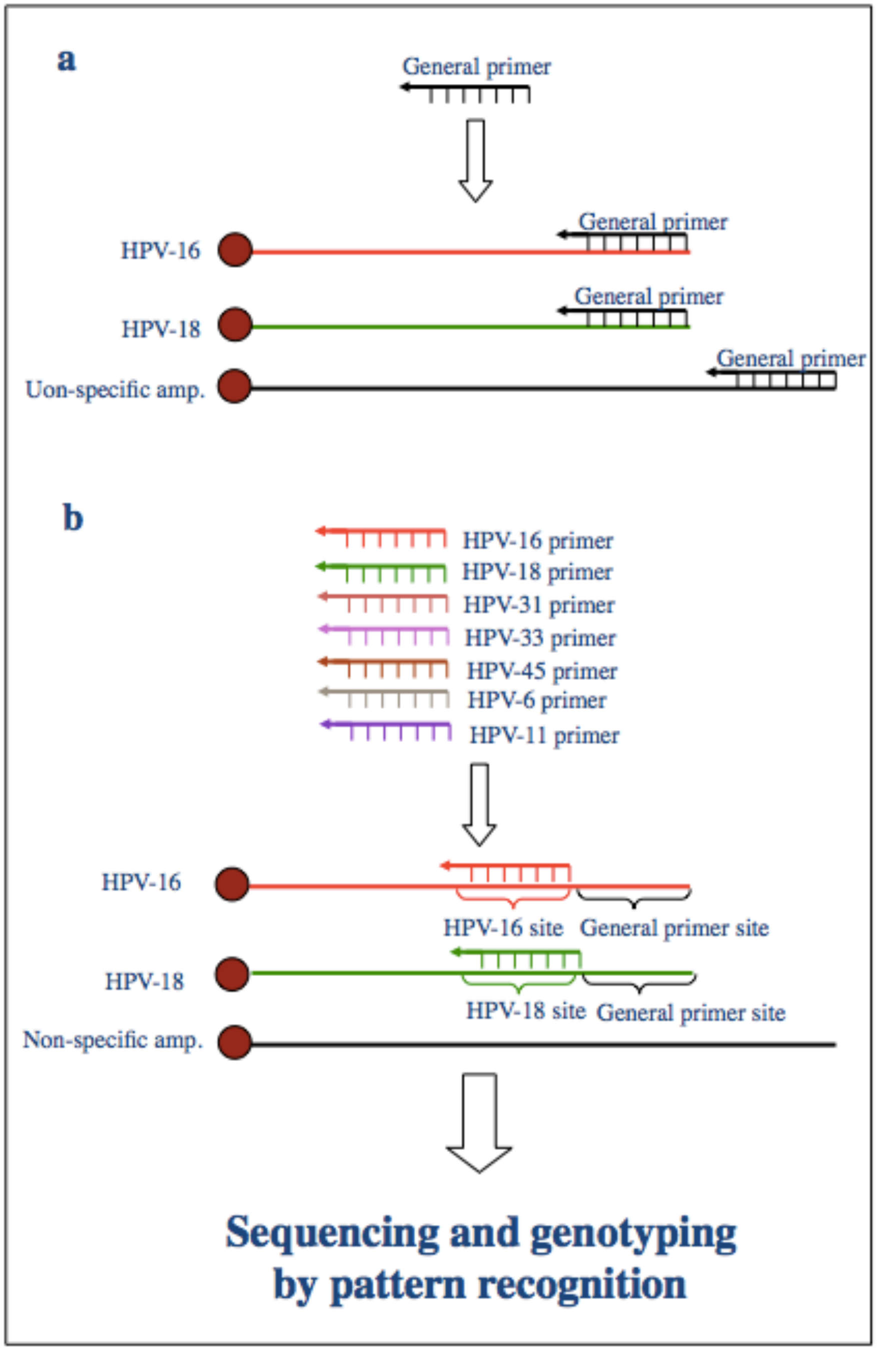
Ronaghi M. Pyrosequencing sheds light on DNA sequencing. Genome Res 2001

Using pyrosequencing to detect HPV infection

- HPV is the most common sexually transmitted infection (STI). HPV is a different virus than HIV and HSV (herpes).
- In most cases, HPV goes away on its own and does not cause any health problems. But when HPV does not go away, it can cause health problems like genital warts and cancer.
- HPV can cause cervical and other cancers including cancer of the vulva, vagina, penis, or anus. It can also cause cancer in the back of the throat, including the base of the tongue and tonsils (called oropharyngeal cancer).



Multiple infections and Non-specific amplification products



Gharizadeh et al. Mol Cell Probes. 2003
Gharizadeh et al. J Mol Diagn. 2005
Gharizadeh et al. Mol Cell Probes. 2006

Using pyroseuencing to detect HPV infection

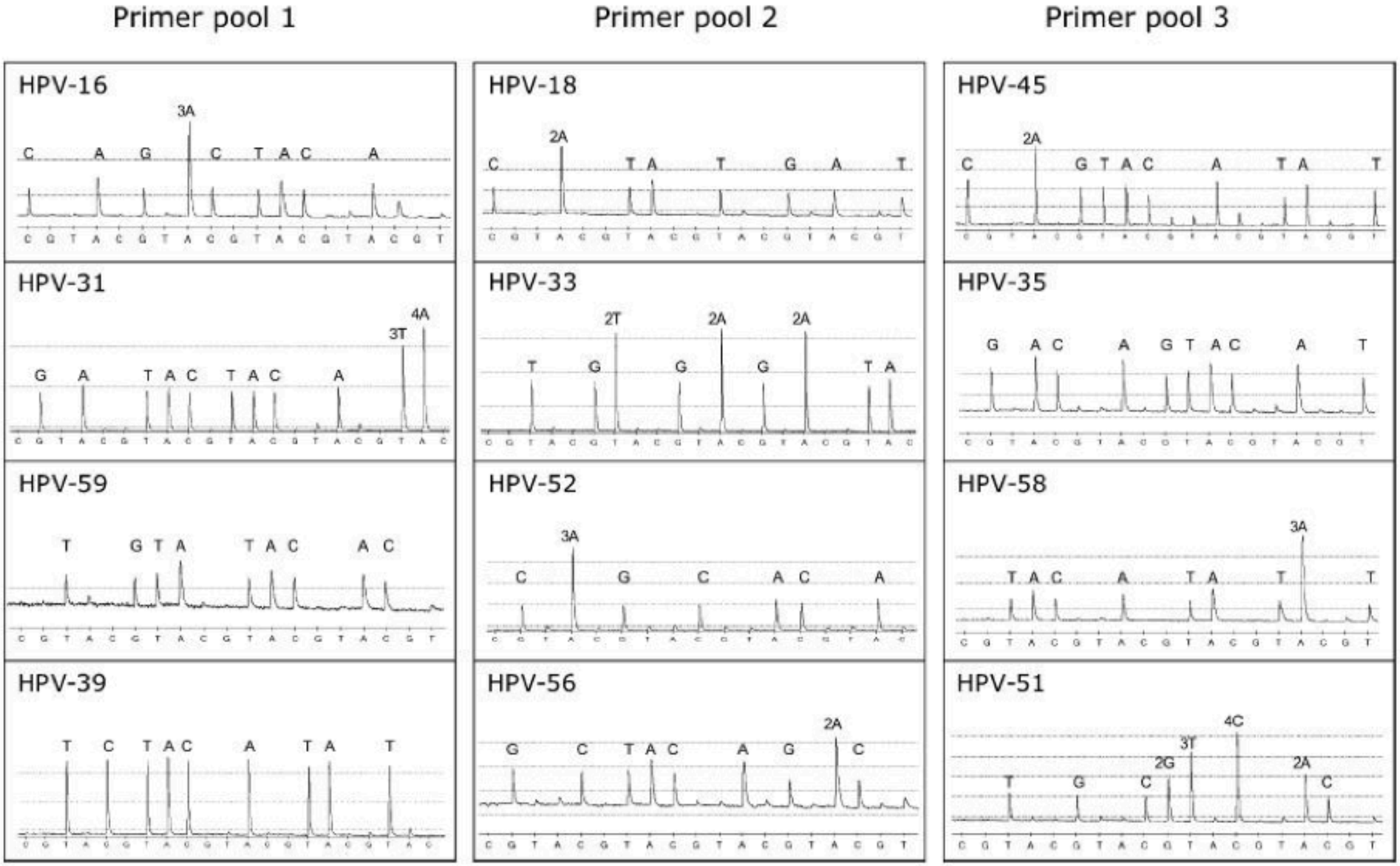
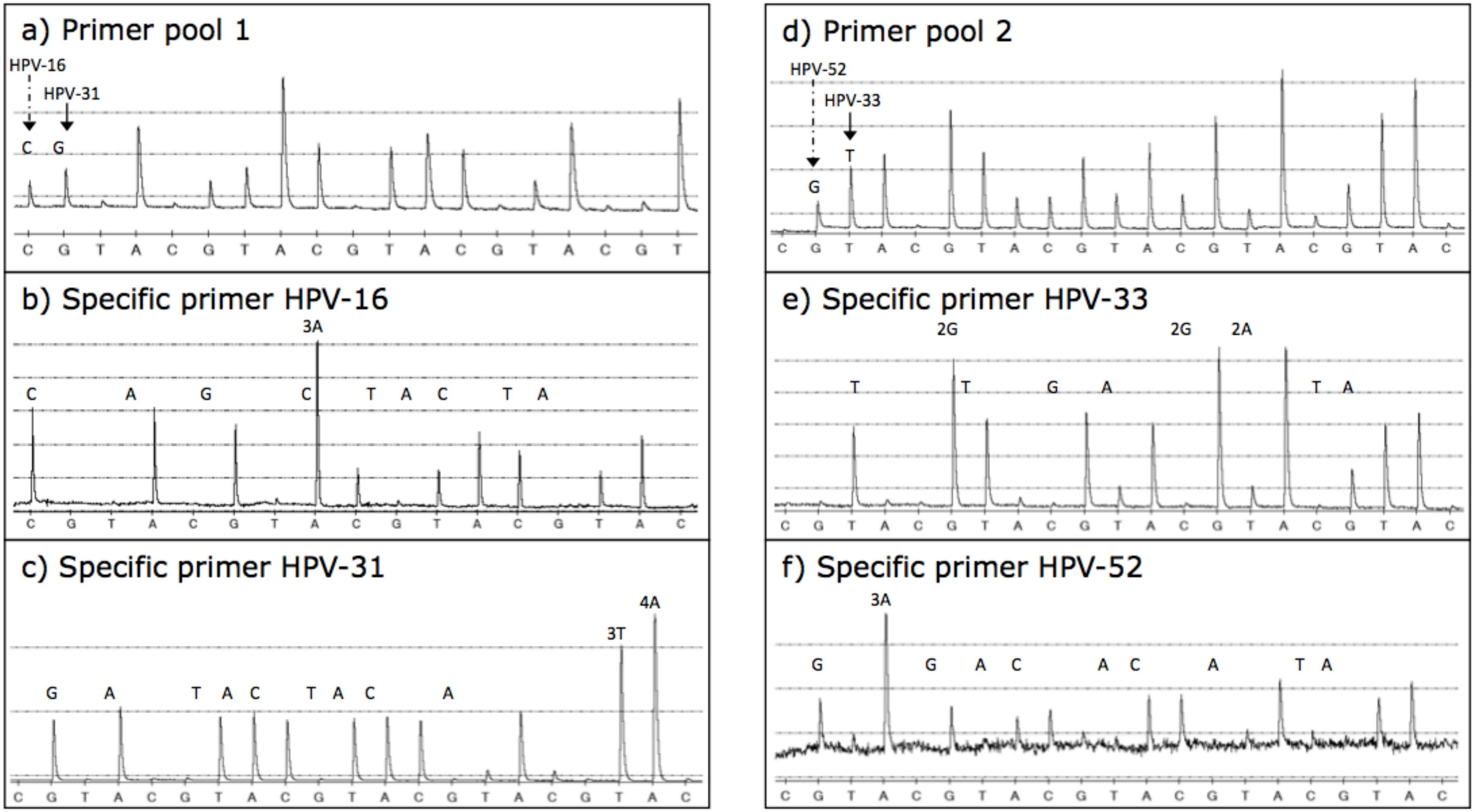


Figure from Gharizadeh, B. et al. "Sentinel-base DNA genotyping using multiple sequencing primers for high-risk human papillomaviruses." Molecular and Cellular Probes, 2006.

HPV co-infection



a) Primer pool 1

b) Specific primer HPV-16

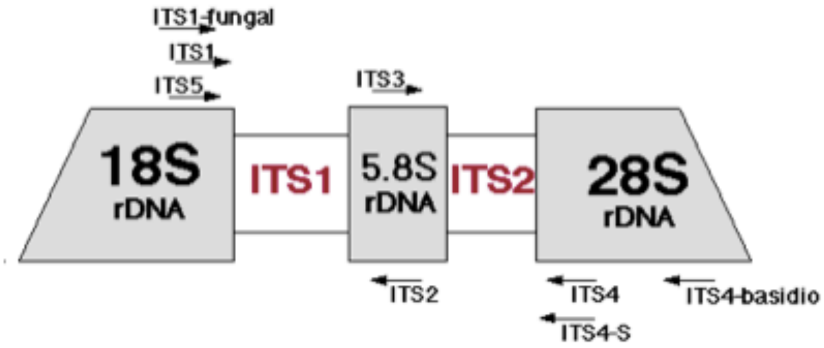
c) Specific primer HPV-31

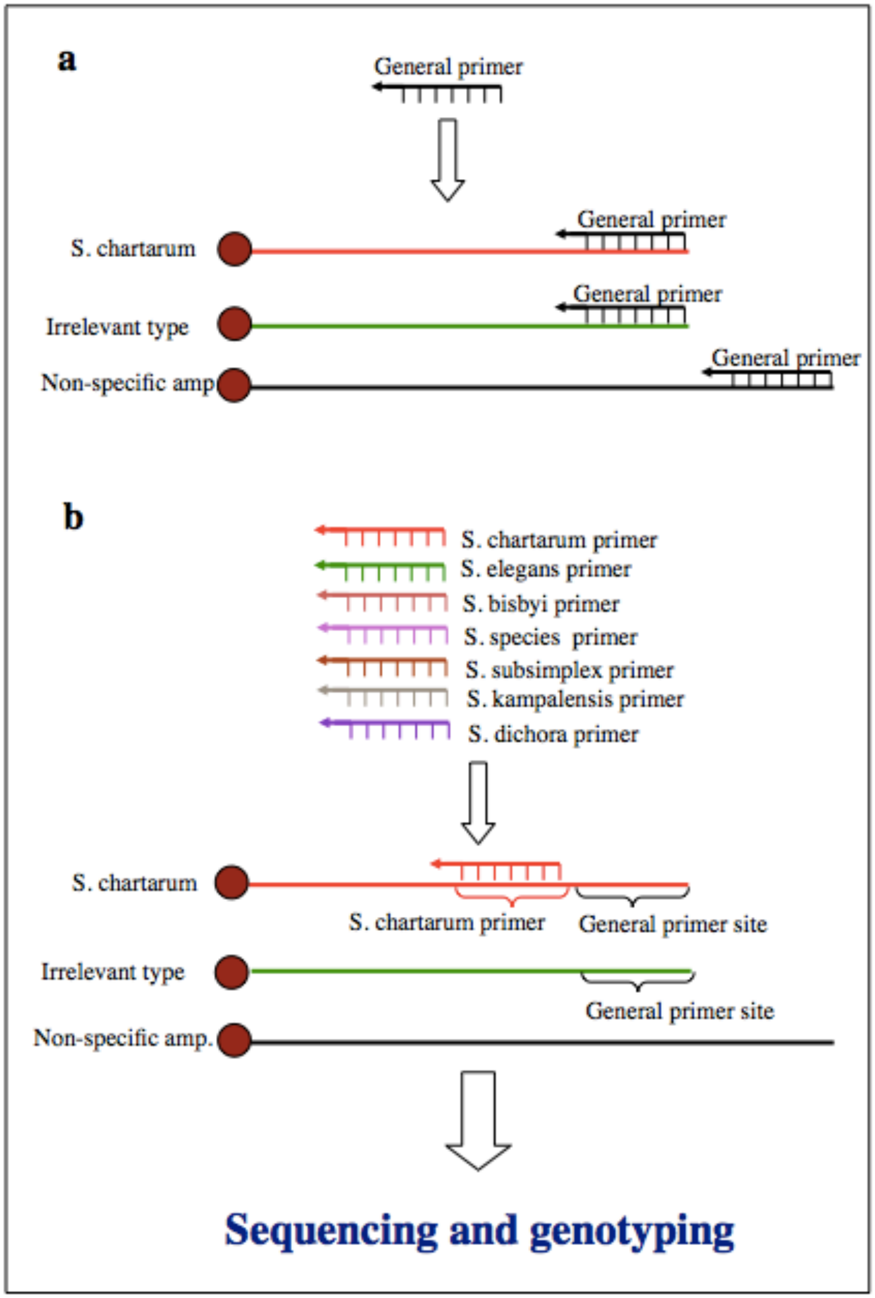
d) Primer pool 2

e) Specific primer HPV-33

f) Specific primer HPV-52

Pyrosequencing for Identification of Fungi





Kaller et al. Submitted 2007

Pyrosequencing for Antibiotic resistance

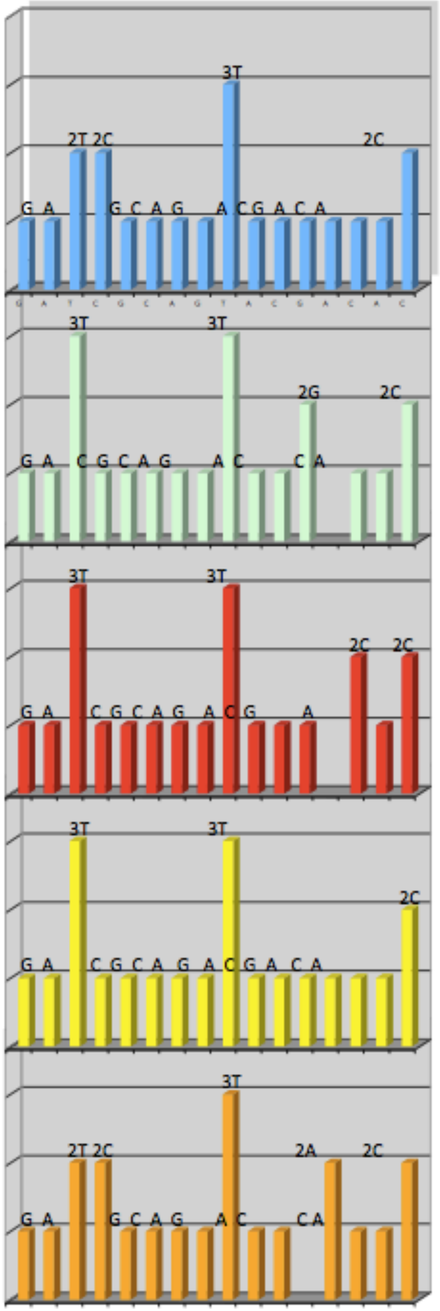
a) **Wild Type:**
GATTCCGCAGTTTACGACACC

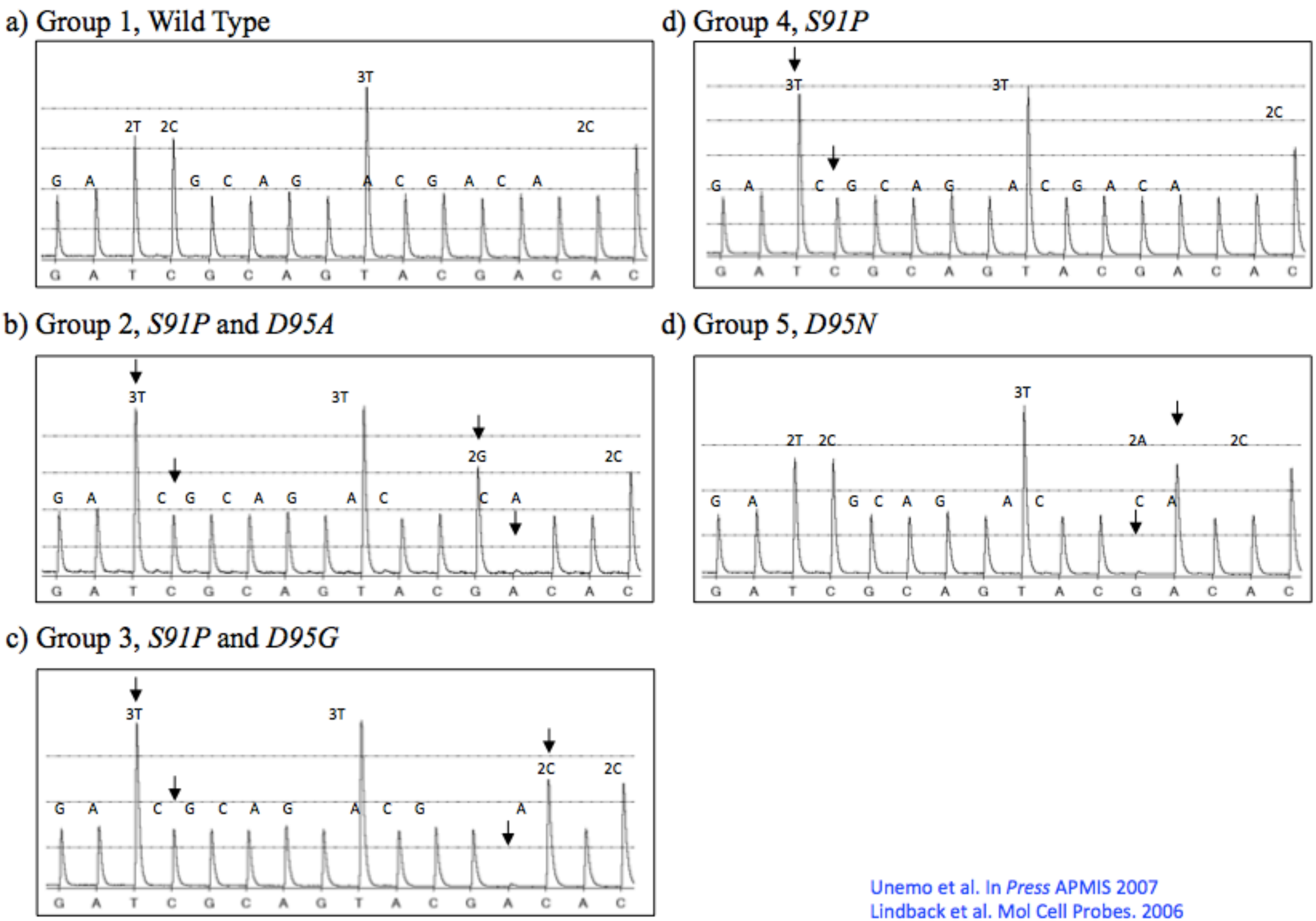
b) **S91P and D95A:**
GATTTCGCAGTTTACGGCACC

c) **S91P and D95G:**
GATTTCGCAGTTTACGCCACC

d) **S91P:**
GATTTCGCAGTTTACGACACC

e) **D95N:**
GATTCCGCAGTTTACAACACC




a) Group 1, Wild Type

d) Group 4, *S91P*

b) Group 2, *S91P* and *D95A*

d) Group 5, *D95N*

c) Group 3, *S91P* and *D95G*

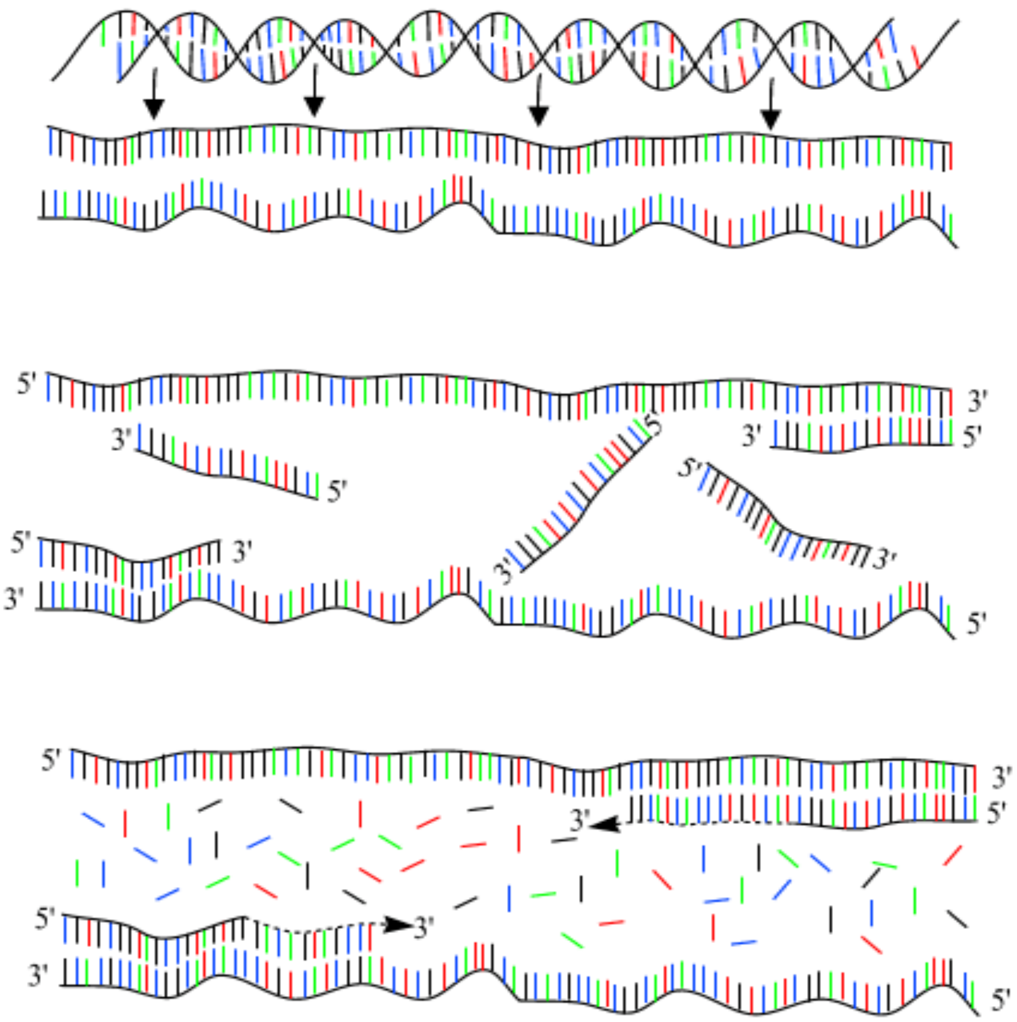
Unemo et al. In *Press* APMIS 2007
Lindback et al. Mol Cell Probes. 2006
Gharizadeh et al. Int J Antimicrob Agents 2005

Pyrosequencing

- Primer design is an important step prior to performing pyrosequencing
  1. For PCR (polymerase chain reaction) amplification of the target
  2. For sequencing the target

PCR : Polymerase Chain Reaction

30 - 40 cycles of 3 steps :



Step 1 : denaturation

1 minut 94 °C

Step 2 : annealing

45 seconds 54 °C

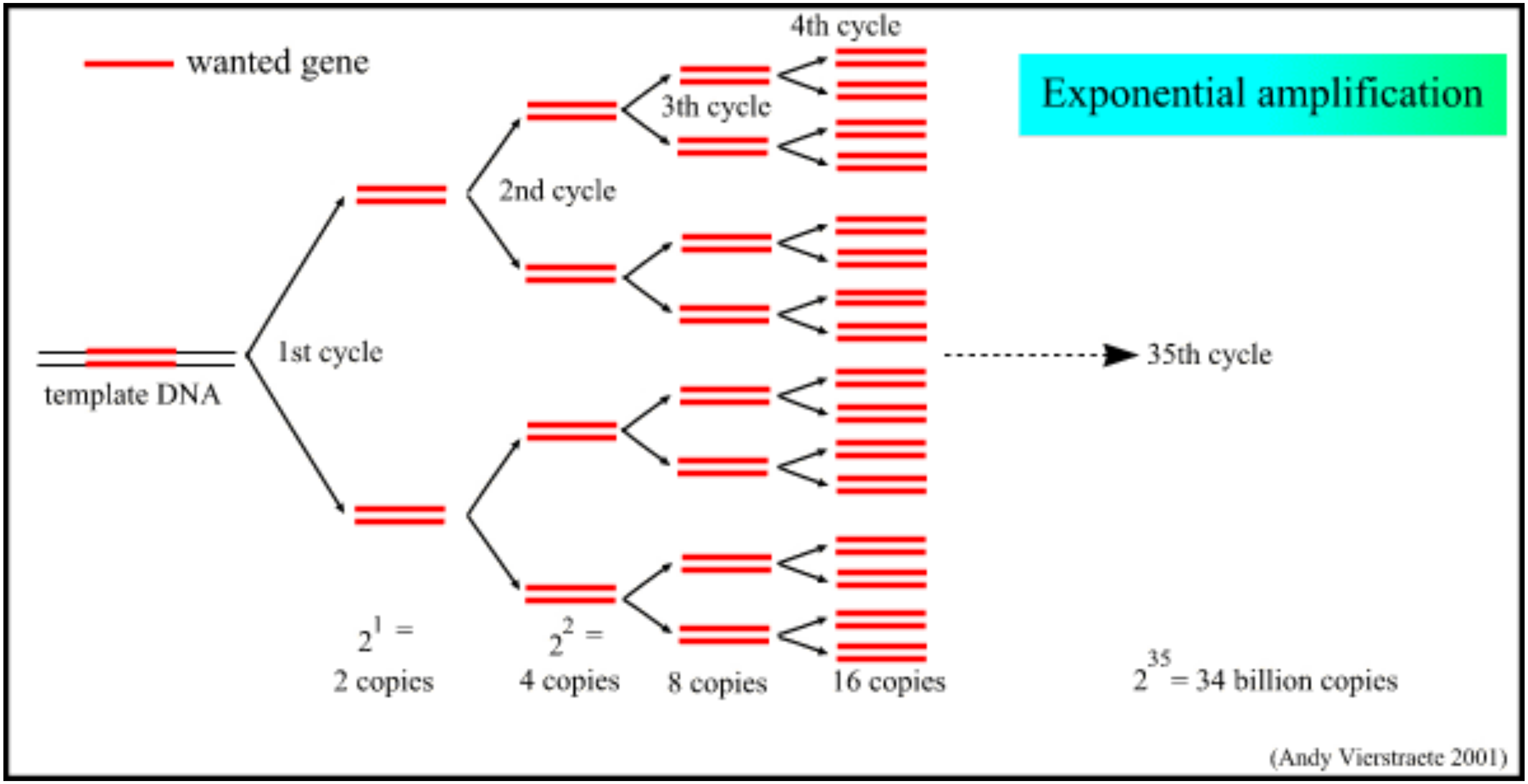
forward and reverse primers !!!

Step 3 : extension

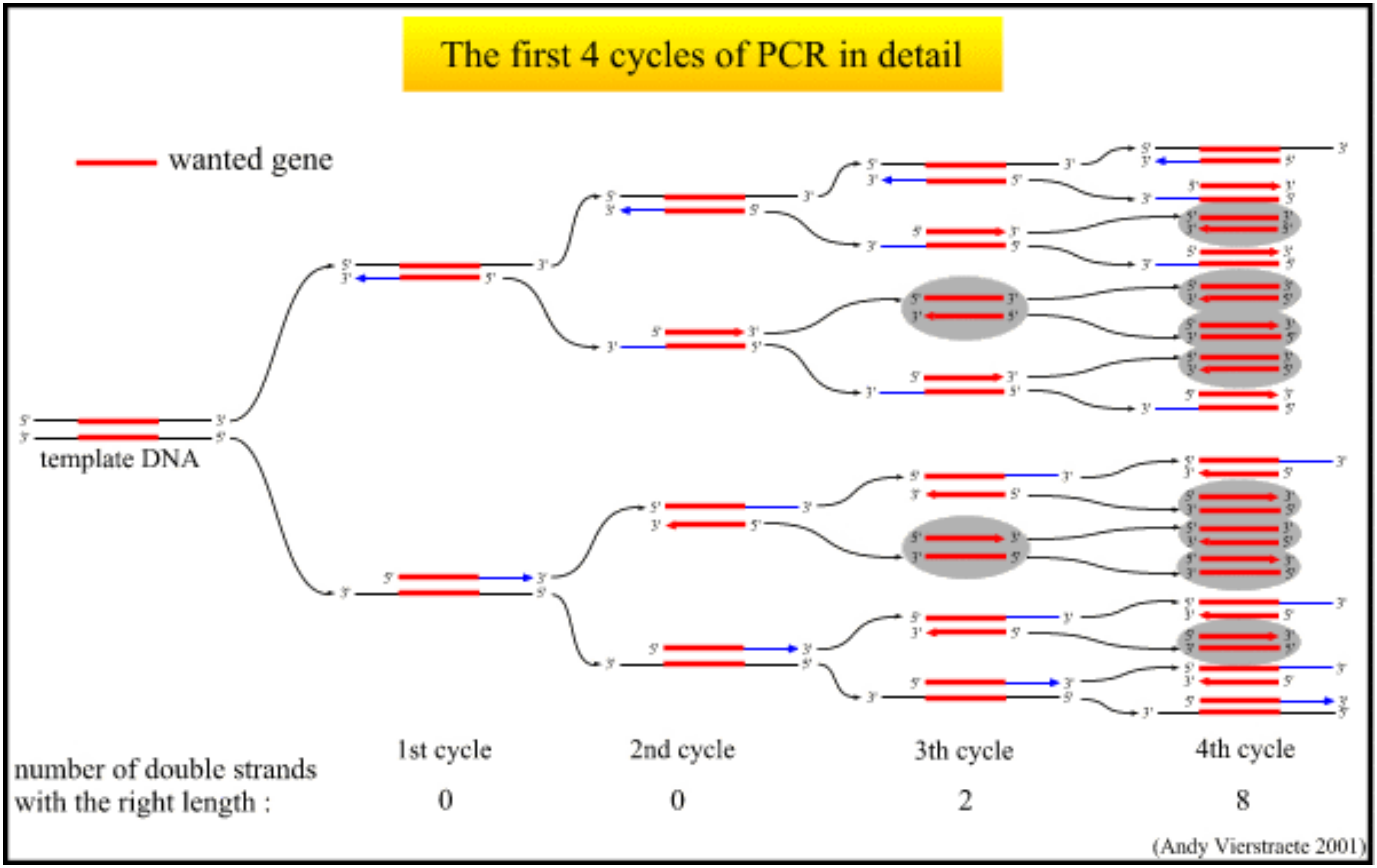
2 minutes 72 °C
only dNTP's

(Andy Vierstraete 1999)

PCR


wanted gene
4th cycle
Exponential amplification
3th cycle
2nd cycle
1st cycle
template DNA
35th cycle
2^1 =
2 copies
2^2 =
4 copies
8 copies
16 copies
2^{35} = 34 billion copies
(Andy Vierstraete 2001)

PCR



(Andy Vierstraete 2001)

Good Primer’s Characteristic

- A melting temperature (Tm) in the range of 52 ^{o}C to 65 ^{o}C
- Absence of dimerization capability
- Absence of significant hairpin formation (>3 bp)
- Lack of secondary priming sites
- Low specific binding at the 3' end (ie. lower GC content to avoid mispriming)

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

Template DNA

```
5'...TCAACTTAGCATGATCGGGTA...GTAGCAGTTGACTGTACAACTCAGCAA...
3'     TGCTAAGTTG                CAGTCAACTGCTAC  TGCTGAGTTG
```

Primer candidate 1 5'-TGCTAAGTTG-3' NOT UNIQUE!

Primer candidate 2 5'-CAGTCAACTGCTAC-3' UNIQUE!

Length

- Primer length has effects on uniqueness and melting/annealing temperature. Roughly speaking, the longer the primer, the more chance that it's unique; the longer the primer, the higher melting/annealing temperature.
- Generally speaking, 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.

Base Composition

Base composition affects hybridization specificity and melting/annealing temperature.

- Random base composition is preferred. We shall avoid long (A+T) and (G+C) rich region if possible.

- Usually, average (G+C) content around 50-60% will give us the right melting/annealing temperature for ordinary PCR reactions, and will give appropriate hybridization stability. However, melting/annealing temperature and hybridization stability are affected by other factors, which we'll discuss later. Therefore, (G+C) content is allowed to change.

Melting Temperature

Melting Temperature, Tm – the temperature at which half the DNA strands are single stranded and half are double-stranded.. Tm is characteristics of the DNA composition; Higher G+C content DNA has a higher Tm due to more H bonds.

Calculation

Shorter than 13: Tm= (wA+xT) * 2 + (yG+zC) * 4

Longer than 13: Tm= 64.9 +41*(yG+zC-16.4)/(wA+xT+yG+zC)

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

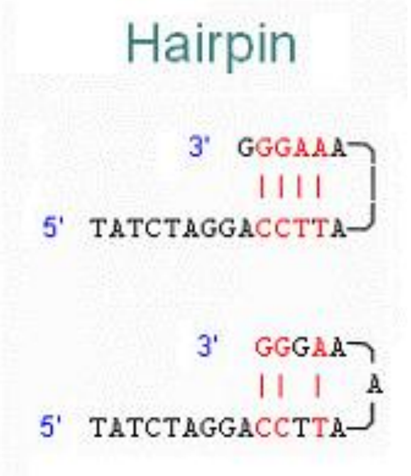
Annealing Temperature

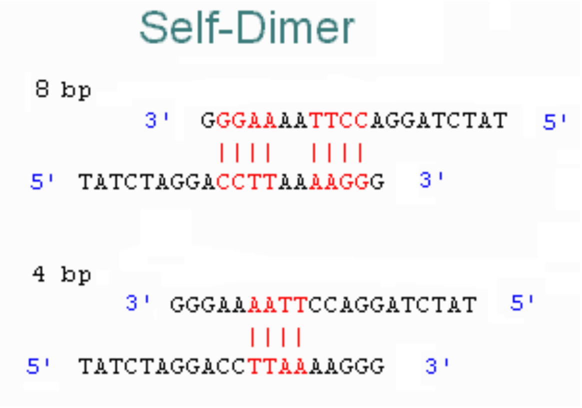
Annealing Temperature, T_{anneal} – the temperature at which primers anneal to the template DNA. It can be calculated from T_m .

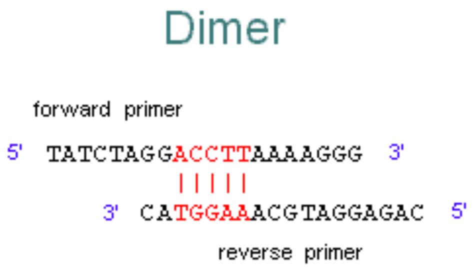
$$T_{anneal} = T_{m\_primer} - 4°C$$

Internal Structure

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




Self-Dimer
8 bp
3' GGGAAAATTCCAGGATCTAT 5'
|||| ||||
5' TATCTAGGACCTTAAAAGGG 3'
4 bp
3' GGGAAAATTCCAGGATCTAT 5'
||||
5' TATCTAGGACCTTAAAAGGG 3'

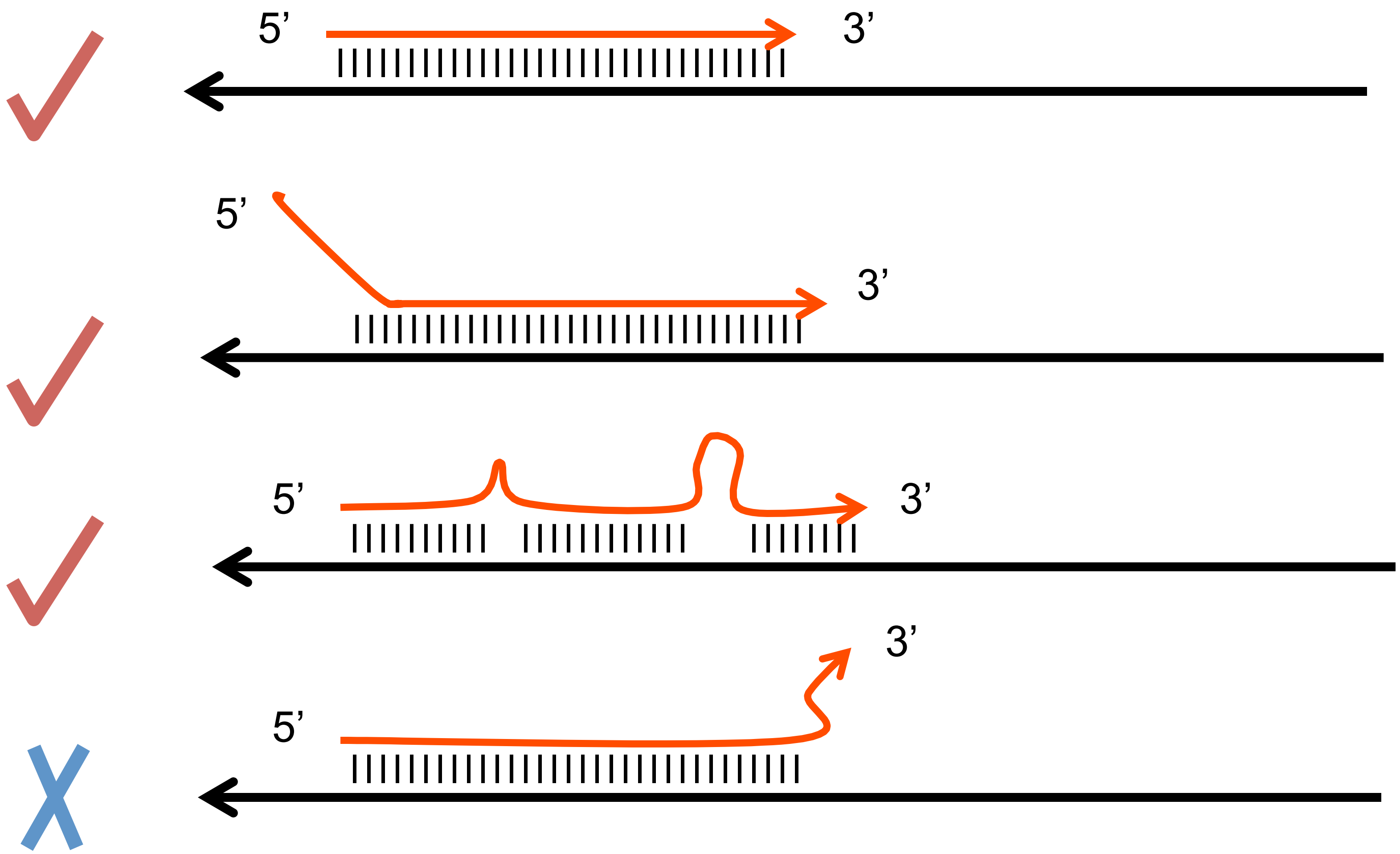


However, sometimes these 2° structures are harmless when the annealing temperature does not allow them to take form. For example, some dimers or hairpins form at 30 °C while during PCR cycle, the lowest temperature only drops to 60 °C.

Primer Pair Matching

- Primers work in pairs – forward primer and reverse primer. Since they are used in the same PCR reaction, it shall be ensured 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_{anneal} are, the better.

Summary ~ when is a “primer” a primer?



Summary ~ Primer Design Criteria

1. Uniqueness: ensure correct priming site;
2. Length: 17-28 bases.This range varies;
3. Base composition: average (G+C) content around 50-60%; avoid long (A+T) and (G+C) rich region if possible;
4. Optimize base pairing: it's critical that the stability at 5' end be high and the stability at 3' end be relatively low to minimize false priming.
5. Melting Tm between 55-80 °C are preferred;
6. Assure that primers at a set have annealing Tm within 2 – 3 °C of each other.
7. Minimize internal secondary structure: hairpins and dimmers shall be avoided.

Computer-Aided Primer Design

Primer design is an **art** when done by human beings, and is **better done by machines**.

Some primer design programs we use:

- **Oligo:** Life Science Software, standalone application
- **GCG**: Accelrys, ICBR maintains the server.
- **Primer3:** MIT, standalone / web application
- **BioTools:** BioTools, Inc. ICBR distributes the license.
- **Others:** GeneFisher, Primer!, Web Primer, NBI oligo program, etc.

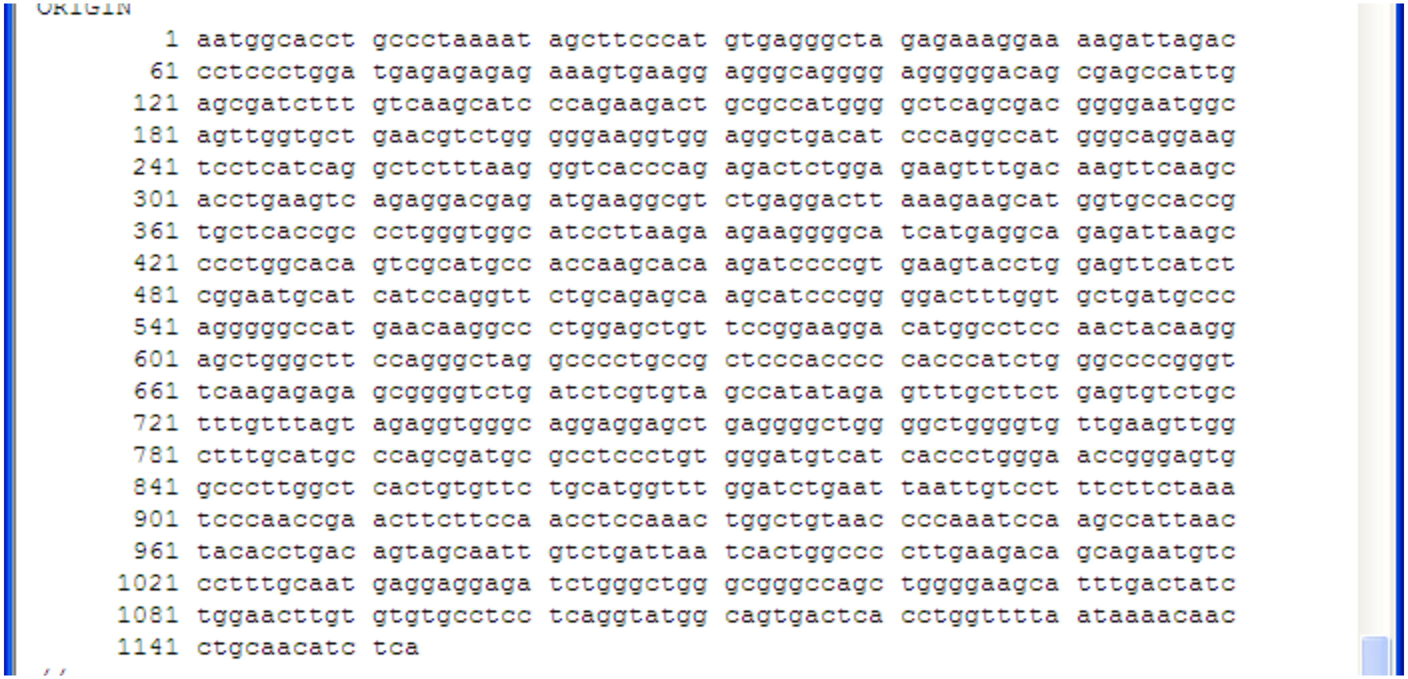
Melting temperature calculation software:

- BioMath: http://www.promega.com/biomath/calc11.htm

Task

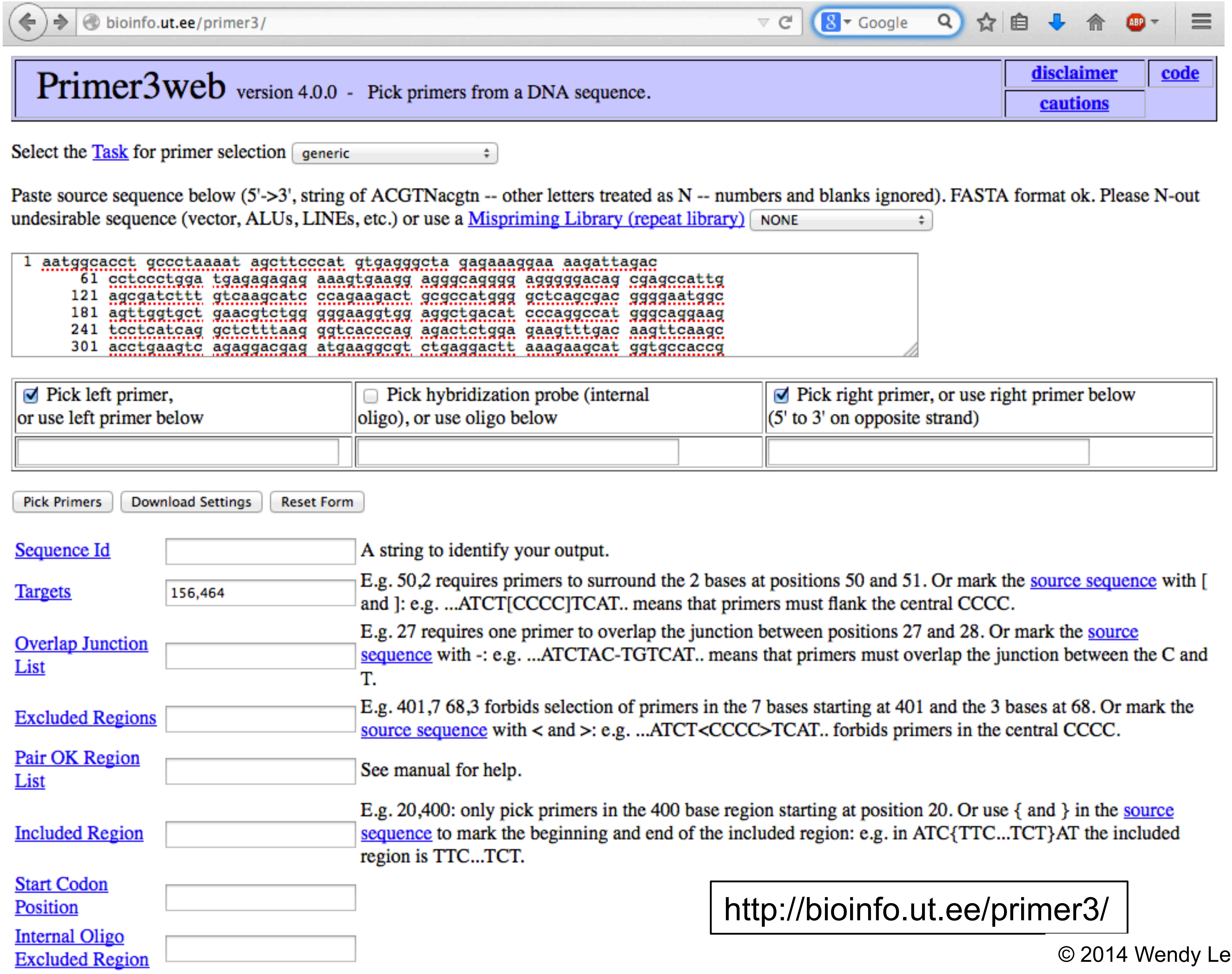
Design a pair of primers for sequence “NM_203378” in NCBI GenBank, so that the coding sequence of human myoglobin will be amplified using PCR reaction.

Between 156..620



```
ORIGIN
        1 aatggcacct gccctaaaat agcttcccat gtgagggcta gagaaaggaa aagattagac
       61 cctccctgga tgagagagag aaagtgaagg agggcagggg agggggacag cgagccattg
      121 agcgatcttt gtcaagcatc ccagaagact gcgccatggg gctcagcgac ggggaatggc
      181 agttggtgct gaacgtctgg gggaaggtgg aggctgacat cccaggccat gggcaggaag
      241 tcctcatcag gctctttaag ggtcacccag agactctgga gaagtttgac aagttcaagc
      301 acctgaagtc agaggacgag atgaaggcgt ctgaggactt aaagaagcat ggtgccaccg
      361 tgctcaccgc cctgggtggc atccttaaga agaaggggca tcatgaggca gagattaagc
      421 ccctggcaca gtcgcatgcc accaagcaca agatccccgt gaagtacctg gagttcatct
      481 cggaatgcat catccaggtt ctgcagagca agcatcccgg ggactttggt gctgatgccc
      541 agggggccat gaacaaggcc ctggagctgt tccggaagga catggcctcc aactacaagg
      601 agctgggctt ccagggctag gcccctgccg ctcccacccc cacccatctg ggccccgggt
      661 tcaagagaga gcggggtctg atctcgtgta gccatataga gtttgcttct gagtgtctgc
      721 tttgtttagt agaggtgggc aggaggagct gaggggctgg ggctggggtg ttgaagttgg
      781 ctttgcatgc ccagcgatgc gcctccctgt gggatgtcat caccctggga accgggagtg
      841 gcccttggct cactgtgttc tgcatggttt ggatctgaat taattgtcct ttcttctaaa
      901 tcccaaccga acttcttcca acctccaaac tggctgtaac cccaaatcca agccattaac
      961 tacacctgac agtagcaatt gtctgattaa tcactggccc cttgaagaca gcagaatgtc
     1021 cctttgcaat gaggaggaga tctgggctgg gcgggccagc tggggaagca tttgactatc
     1081 tggaacttgt gtgtgcctcc tcaggtatgg cagtgactca cctggtttta ataaaacaac
     1141 ctgcaacatc tca
//
```

© 2014 Wendy Lee



Primer3web version 4.0.0 - Pick primers from a DNA sequence.

Select the Task for primer selection: generic

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): NONE

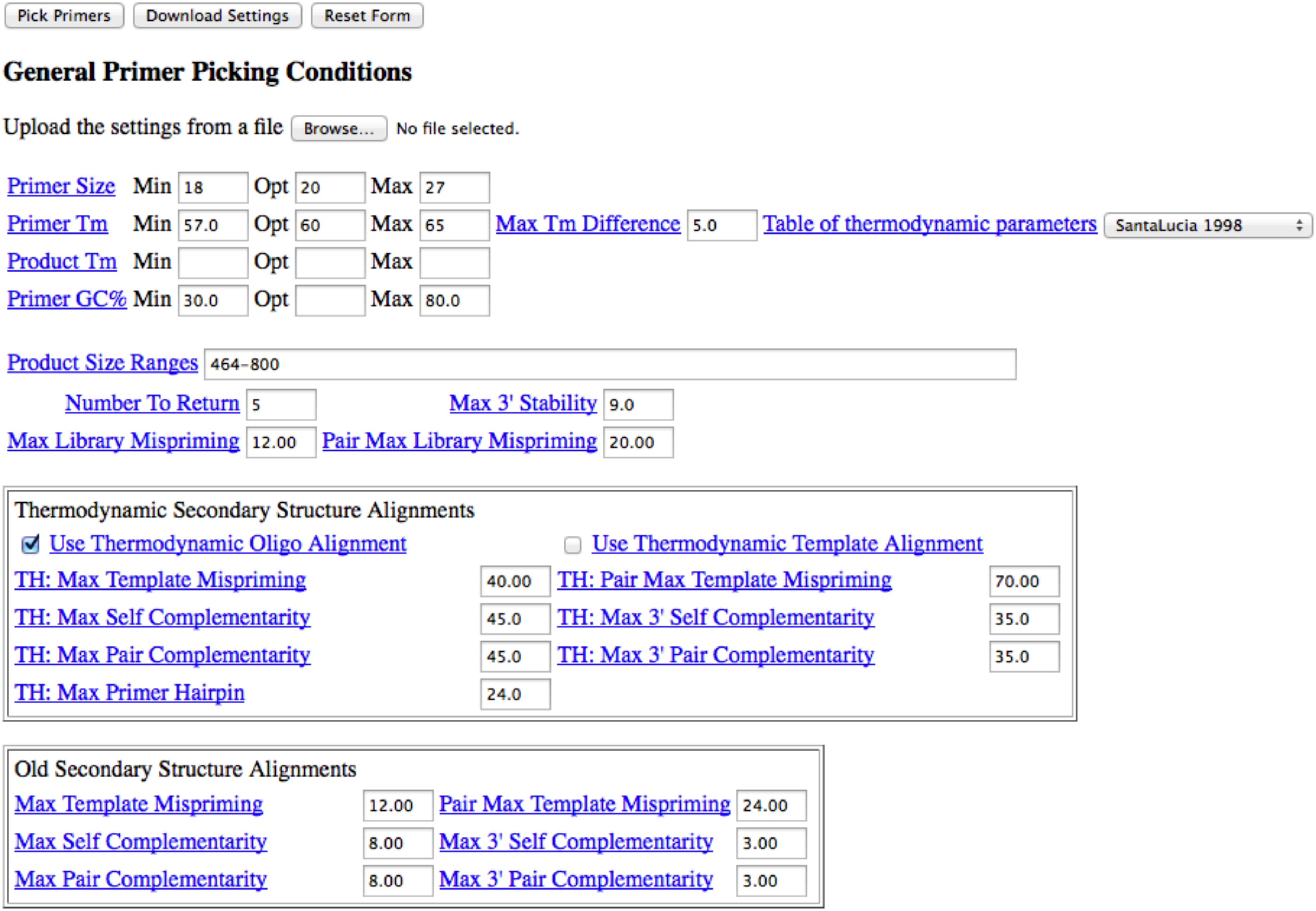
```
1 aatggcacct gccctaaaat agcttcccat gtgagggcta gagaaaggaa aagattagac
      61 cctccctgga tgagagagag aaagtgaagg agggcagggg agggggacag cgagccattg
     121 agcgatcttt gtcaagcatc ccagaagact gcgccatggg gctcagcgac ggggaatggc
     181 agttggtgct gaacgtctgg gggaaggtgg aggctgacat cccaggccat gggcaggaag
     241 tcctcatcag gctctttaag ggtcacccag agactctgga gaagtttgac aagttcaagc
     301 acctgaagtc agaggacgag atgaaggcgt ctgaggactt aaagaagcat ggtgccaccg
```

☑ Pick left primer, or use left primer below	□ Pick hybridization probe (internal oligo), or use oligo below	☑ Pick right primer, or use right primer below (5' to 3' on opposite strand)

Pick Primers | Download Settings | Reset Form

Sequence Id		A string to identify your output.
Targets	156,464	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.
Overlap Junction List		E.g. 27 requires one primer to overlap the junction between positions 27 and 28. Or mark the source sequence with -: e.g. ...ATCTAC-TGTCAT.. means that primers must overlap the junction between the C and T.
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.
Pair OK Region List		See manual for help.
Included Region		E.g. 20,400: only pick primers in the 400 base region starting at position 20. Or use { and } in the source sequence to mark the beginning and end of the included region: e.g. in ATC{TTC...TCT}AT the included region is TTC...TCT.
Start Codon Position		
Internal Oligo Excluded Region		

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



General Primer Picking Conditions

Upload the settings from a file Browse... No file selected.

Primer Size	Min	18	Opt	20	Max	27
Primer Tm	Min	57.0	Opt	60	Max	65
Product Tm	Min		Opt		Max	
Primer GC%	Min	30.0	Opt		Max	80.0

Max Tm Difference 5.0 Table of thermodynamic parameters SantaLucia 1998

Product Size Ranges 464-800

Number To Return	5	Max 3' Stability	9.0
Max Library Mispriming	12.00	Pair Max Library Mispriming	20.00

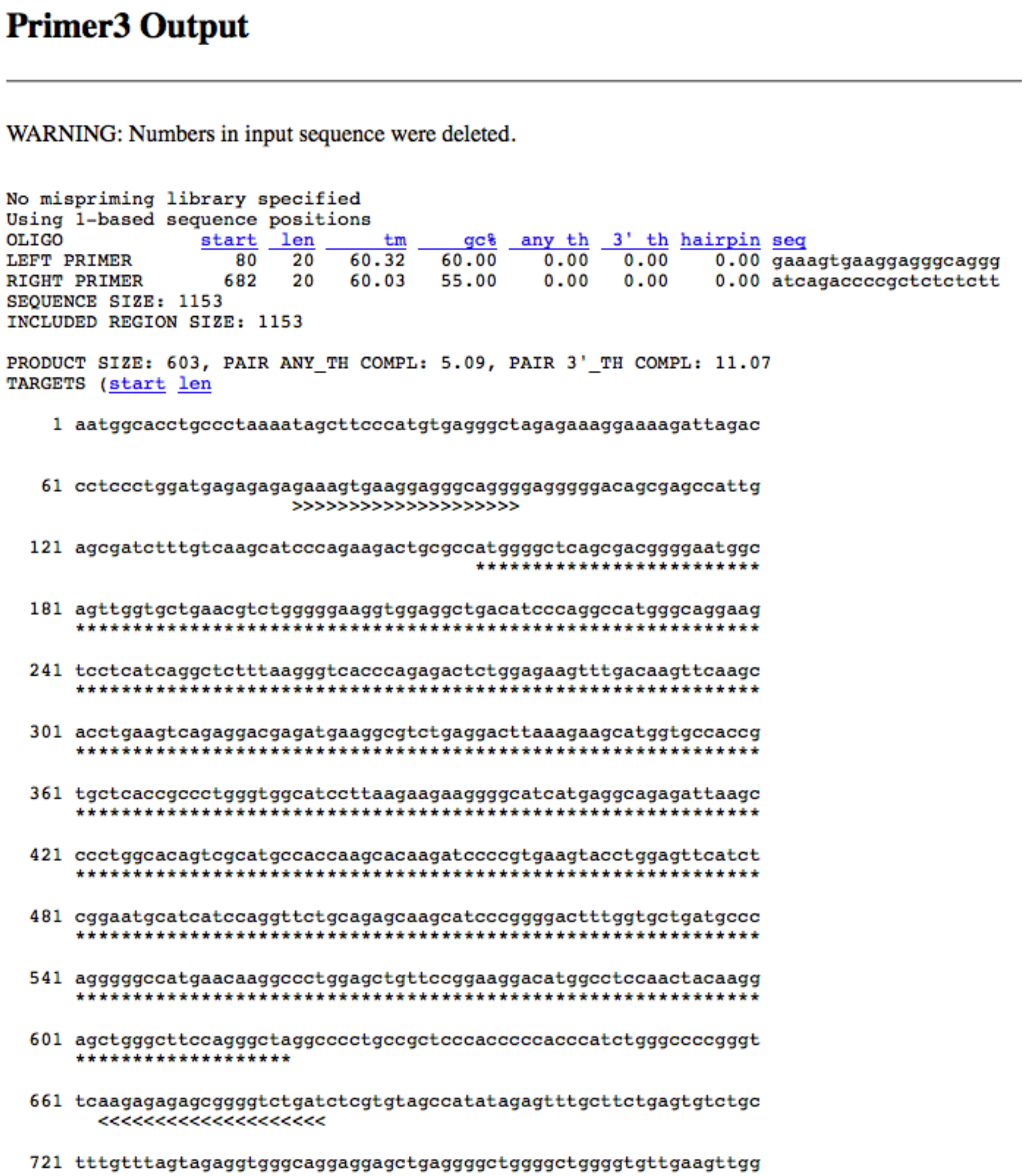
Thermodynamic Secondary Structure Alignments

☑ Use Thermodynamic Oligo Alignment ☐ Use Thermodynamic Template Alignment

TH: Max Template Mispriming	40.00	TH: Pair Max Template Mispriming	70.00
TH: Max Self Complementarity	45.0	TH: Max 3' Self Complementarity	35.0
TH: Max Pair Complementarity	45.0	TH: Max 3' Pair Complementarity	35.0
TH: Max Primer Hairpin	24.0		

Old Secondary Structure Alignments

Max Template Mispriming	12.00	Pair Max Template Mispriming	24.00
Max Self Complementarity	8.00	Max 3' Self Complementarity	3.00
Max Pair Complementarity	8.00	Max 3' Pair Complementarity	3.00



Primer3 Output

WARNING: Numbers in input sequence were deleted.

```
No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len      tm     gc%  any th  3' th hairpin seq
LEFT PRIMER         80   20   60.32   60.00    0.00   0.00    0.00 gaaagtgaaggagggcaggg
RIGHT PRIMER       682   20   60.03   55.00    0.00   0.00    0.00 atcagaccccgctctctctt
SEQUENCE SIZE: 1153
INCLUDED REGION SIZE: 1153

PRODUCT SIZE: 603, PAIR ANY_TH COMPL: 5.09, PAIR 3'_TH COMPL: 11.07
TARGETS (start len

    1 aatggcacctgccctaaaatagcttcccatgtgagggctagagaaaggaaaagattagac


   61 cctccctggatgagagagagaaagtgaaggagggcaggggaggggacagcgagccattg
                         >>>>>>>>>>>>>>>>>>>>

  121 agcgatctttgtcaagcatcccagaagactgcgccatggggctcagcgacggggaatggc
                                         *************************

  181 agttggtgctgaacgtctggggaaggtggaggctgacatcccaggccatgggcaggaag
      ************************************************************

  241 tcctcatcaggctctttaagggtcacccagagactctggagaagtttgacaagttcaagc
      ************************************************************

  301 acctgaagtcagaggacgagatgaaggcgtctgaggacttaaagaagcatggtgccaccg
      ************************************************************

  361 tgctcaccgccctgggtggcatccttaagaagaaggggcatcatgaggcagagattaagc
      ************************************************************

  421 ccctggcacagtcgcatgccaccaagcacaagatccccgtgaagtacctggagttcatct
      ************************************************************

  481 cggaatgcatcatccaggttctgcagagcaagcatcccggggactttggtgctgatgccc
      ************************************************************

  541 agggggccatgaacaaggccctggagctgttccggaaggacatggcctccaactacaagg
      ************************************************************

  601 agctgggcttccagggctaggcccctgccgctcccacccccacccatctgggccccgggt
      *******************

  661 tcaagagagagcggggtctgatctcgtgtagccatatagagtttgcttctgagtgtctgc
        <<<<<<<<<<<<<<<<<<<<

  721 tttgtttagtagaggtgggcaggaggagctgaggggctggggctggggtgttgaagttgg
```

Multiplex PCR

- Multiple primer pairs can be added in the same tube to do the PCR
- Good for amplifying multiple sites
- Application example: genome identification
- Design difficulty
  - Melting temperatures should be similar
  - No dimer formulation

Universal Primers

Primers can be designed to amplify only one product.

Primers can also be designed to amplify multiple products. We call such primers “universal primers”. For example, design primers to amplify all HPV genes.

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 primer at the 5’ conservative region.
4. Design reverse primer at the 3’ conservative regions.
5. Matching forward and reverse primers to find the best pair.
6. Ensure uniqueness in all template sequences.
7. Ensure uniqueness in possible contaminant sources.

Semi-Universal Primers

Primers can be designed to amplify only a subset of template sequences from a large group of similar sequences. For example, design primer to amplify HPV type 1 and type 6 gene, but not other types.

Strategy:

1. Align all types of HPV genes.
2. Identify a subset of genes that are more similar to each other than to other subsets. In this case, type 1 and type 6.
3. Find the 5’ and 3’ regions that are conserved between type 1 and type 6, but are variable in other types.
4. Design forward primers from the 5’ region and reverse primers from the 3’ region.
5. Matching forward and reverse primers to find the best pair.
6. Ensure uniqueness in all template sequences.
7. Ensure uniqueness in possible contaminant sources.

Guessmer

- In some cases, DNA sequences are either unavailable or difficult to align. Then, a single/group of related proteins can be back translated into nucleotide sequences that will be used as template to design primers/probes. We call such primers “guessmer”.

- Back translation is both problematic and feasible. While the genetic codes are degenerate, different organisms do show preferential biases in codon usage, which can be used to limit the possible back-translated nucleotide sequences.

Guessmer

Strategy:

- Back translate the protein sequence using corresponding codon usage table. Identify 5’ and 3’ regions where there is the least ambiguity.
- Design and match forward and reverse primers as before. But the primers shall be about 30 bases long in order to offset the decreased hybridization specificity caused by mismatched bases.
- Set higher annealing temperature to increase the primer annealing stringency.

Summary ~ Advanced Primer Design

Primers can be designed to serve various purposes. Universal primer, semi-universal primer, guessmers are some of them. There are many more fields where primer design skills are required, such as real-time PCR, population polymorphism study (microsatellite, AFLP, SNP …), internal probe design, and so on.

However, the basic rules always apply –
achieve the appropriate hybridization specificity and stability.