Deconvolving Sequence Variations in Mixed DNA Populations

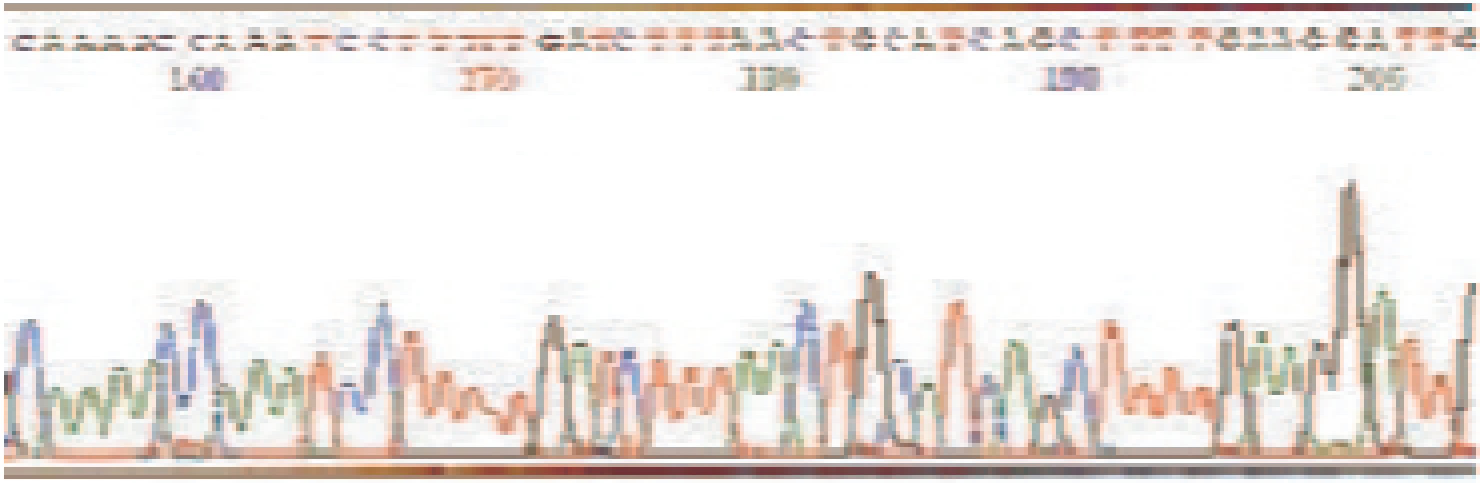
Andy Wildenberg, Steven Skiena,
Pavel Sumazin
Department of Computer Science
SUNY Stony Brook

Overview

- Motivation
- Problem Definition
- Theoretical Results
- Experimental Results
- SNPs
- Future Directions

Gel electrophorisis sequencing

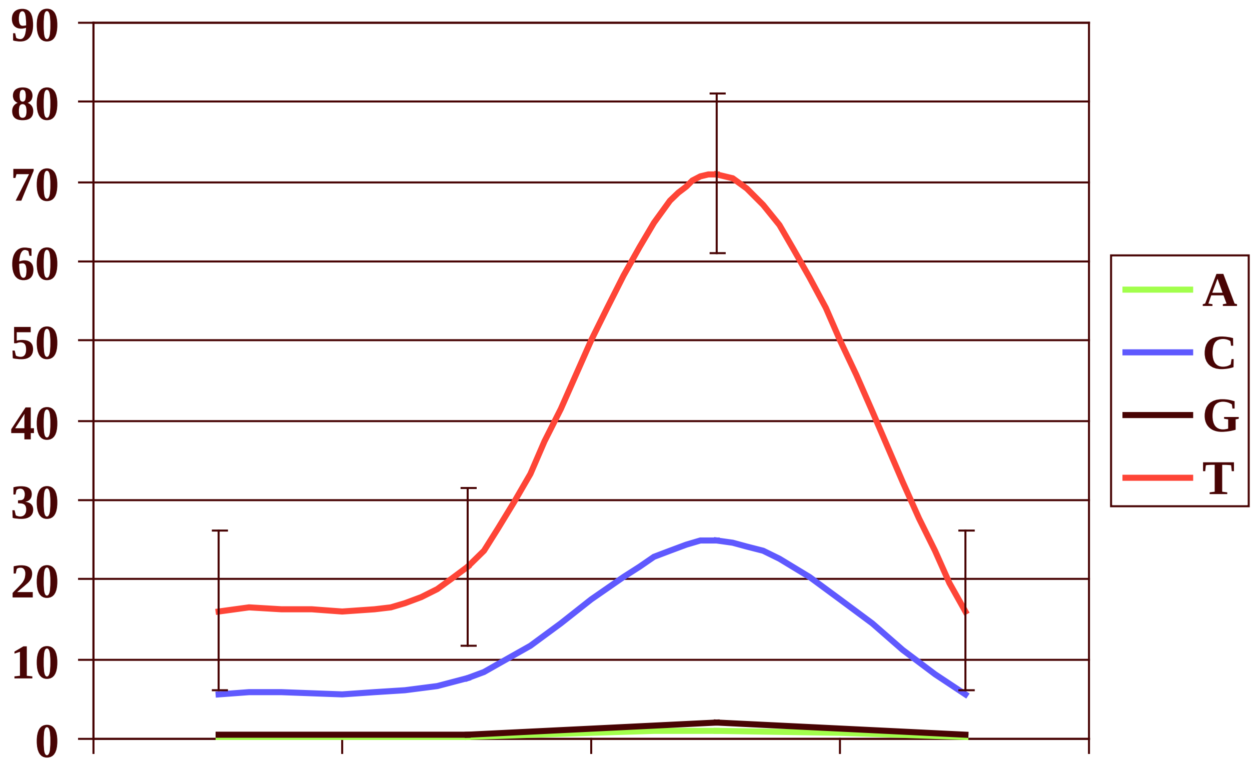
- homogeneous DNA sample
- four output traces
- largest peak defines underlying sequence
- likelihood of correct call



More accurate sequencing

By using advanced single-photon detectors and other technologies, BioPhotonics has the capability to not only detect but accurately determine the relative frequency of each base at each position to within 10%, and expects to reduce this error rate in the near future.

Sequencing inhomogeneous data



Basepair 10: A=1% C=25% G=2% T=71%

relative weights may yield info on presence/frequency of mutations

BioPhotonics sequencers

- smaller (8”x8”x16” -- 20 x 20 x 40 cm)
- cheaper ($10k-20k)
- more accurate
- ideal for diagnostic situations (one in every doctor’s office)

Detecting acquired mutations

- individualized medicine
- microarrays can diagnose leukemia and breast cancer subtypes
- Sanger sequencing is more general tool
- must be able to sequence heterogeneous mix if dealing with acquired mutations

Problem Definitions

- Base calling
- Deconvolution
- Population frequency determination

Base calling

- Assume external program provides F(i,j), the percentage of base i observed at position j
- F(i,j) contains errors

Mutation deconvolution

- Input
  - S, a wildtype sequence
  - V, a set of legal variations/mutations
  - Experimental profile

```
TGTTGACTCATCCC     Wildtype
  AACCACTCCT   C   other
          A
```

Mutation deconvolution

- Output
  - – smallest subset V’ ⊆ V such that the mutations cover the experimental profile

Profile

```
TGTTGACTCATCCC    Wildtype
  AACCACTCCT  C   other
          A
```

Solution

```
TGTTGACTCATCCC    Wildtype
tgttgCACTCATccC   Ins(6,C)
tgAACactcatccc    Sub(3,AAC)
tgttgactcaCcc     Del(11,1)
```

Population Frequency Determination

- Input:

  S, a Wildtype sequence

  V, a set of allowable variations

  F(i,j), an observed profile

- Output:

  w_i, a list of weights assigned to each variation so that their sum most closely matches F(i,j).

Theoretical Results

Kinds of Mutations

ACTGTTGACTCATCCC	Wildtype
ACTGTTCACTCATCCC	Substitution - Sub(7,C)
ACTGTTCGATCATCCC	Substitution - Sub(7,CGA)
ACTGTTACTCATCCC	Deletion- Del(7,1)
ACTGTTTGACTCATCCC	Insertion - Ins(7,T)

Some mutation classes are easy to deconvolve

- All SNPs
- All substitutions up to a given length
- Both solved by greedy algorithm, working left to right

Most mutation classes are hard to deconvolve

- All mutations from a list
- All possible deletions
- All possible insertions

- Hard by reduction from Set-Cover
  - hard to solve, hard to approximate

Substitutions from a list
(reduction from set cover)

- Set cover problem

  N={1,2,3,4}, M={{1,2},{2,3},{3,4}}

- Deconvolution problem

AAAA Wildtype
CCCC rest of profile

CCAA -- {1,2}
ACCA -- {2,3}   mutation list
AACC -- {3,4}

Arbitrary Insertion/Deletion

- Construct long wildtype encouraging certain kinds of insertions/deletions, penalizing others
- Insertion reduction example

```
#**--#-**-#--**#----#**--#-**-#--**#----#**--#-**-#--**#----
                                                            #**--#-**-#--**#----
1--**1*--*1**--1****1--**1*--*1**--1****1--**1*--*1**--1****1--**1*--*1**--1****
                1111                1111                1111                1111
```

- Deletion reduction similar

Same length deletions mask each other

Mutation set

TGTTGACTCATCCC	Wildtype
TGTGACTCATCCC	D(4,1)
TGTTGATCATCCC	D(7,1)

Profile

TGTTGACTCATCCC	Wildtype
GACTCATC	other

Insertions may mask each other

Mutation set

TGTTGACTCATCCC	Wildtype
TGTATGACTCATCCC	I(4,A)
TGTTGACTTCATCCC	I(8,T)

Profile

TGTTGACTCATCCC	Wildtype
ATGACTCATCCC	other

Experimental Results

Assumptions

- F(i,j) -- observed frequency of base i at location j
- F(i,j) is corrupted by Uniform noise
- list of all possible mutations is known in advance

Base calling

- Set thresholds t_{hi}, t_{lo}
- C(i,j) =
  - *Present* if $F(i,j) > t_{hi}$
  - *Absent* if $F(i,j) < t_{lo}$
  - *NoCall* if $t_{lo} < F(i,j) < t_{hi}$

Mutation Deconvolution

- Find a minimal set of mutations so that
  - all *Present* are covered
  - no *Absent* are covered
  - all mutations are from the specified list
- A* search (DFS)
- Aggressive pruning

Population Frequency Determination

- Take solution to Deconvolution
- Find weights for the mutations so that they match observed weights F(i,j)

Deconvolution solution as overconstrained linear system

TGTTGACT	Wildtype	
TGTACACT	mutation 1	Sub(4,AC)
TGAAGACT	mutation 2	Sub(3,AA)

$$
\begin{aligned}
F(T,3) &= ww + w1 \\
F(A,3) &= w2 \\
F(T,4) &= ww \\
F(A,4) &= w1 + w2 \\
F(C,5) &= w1 \\
F(G,5) &= ww + w2 \\
F(A,6) &= ww + w1 + w2
\end{aligned}
$$

plus lots of degenerate equations

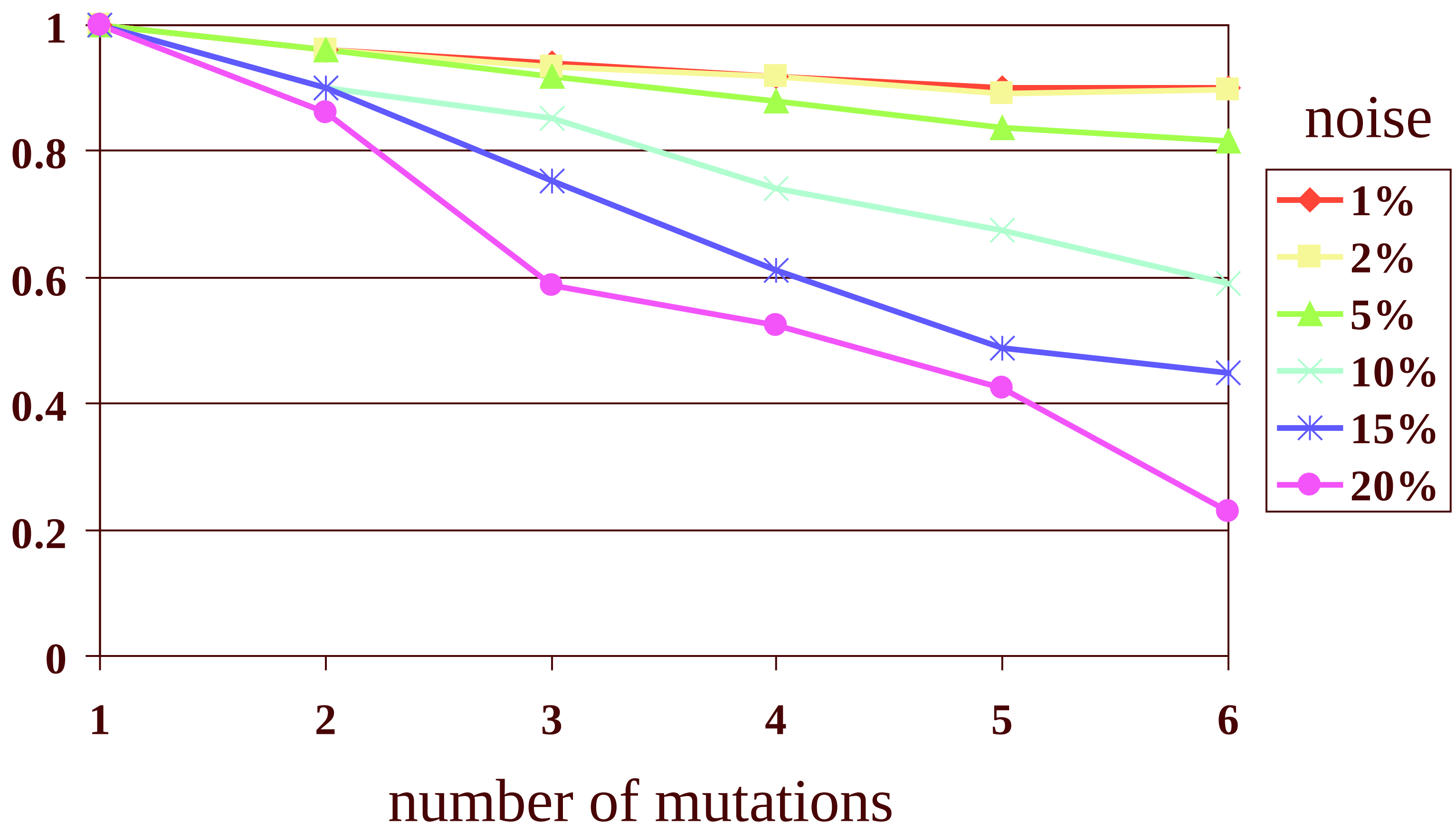
Simulated Results

- p53 Mutation catalog
- International Agency for Research on Cancer, Lyon, France, Version R5 (June 2001)
- 2362 distinct mutations from many sources (14755 reported)

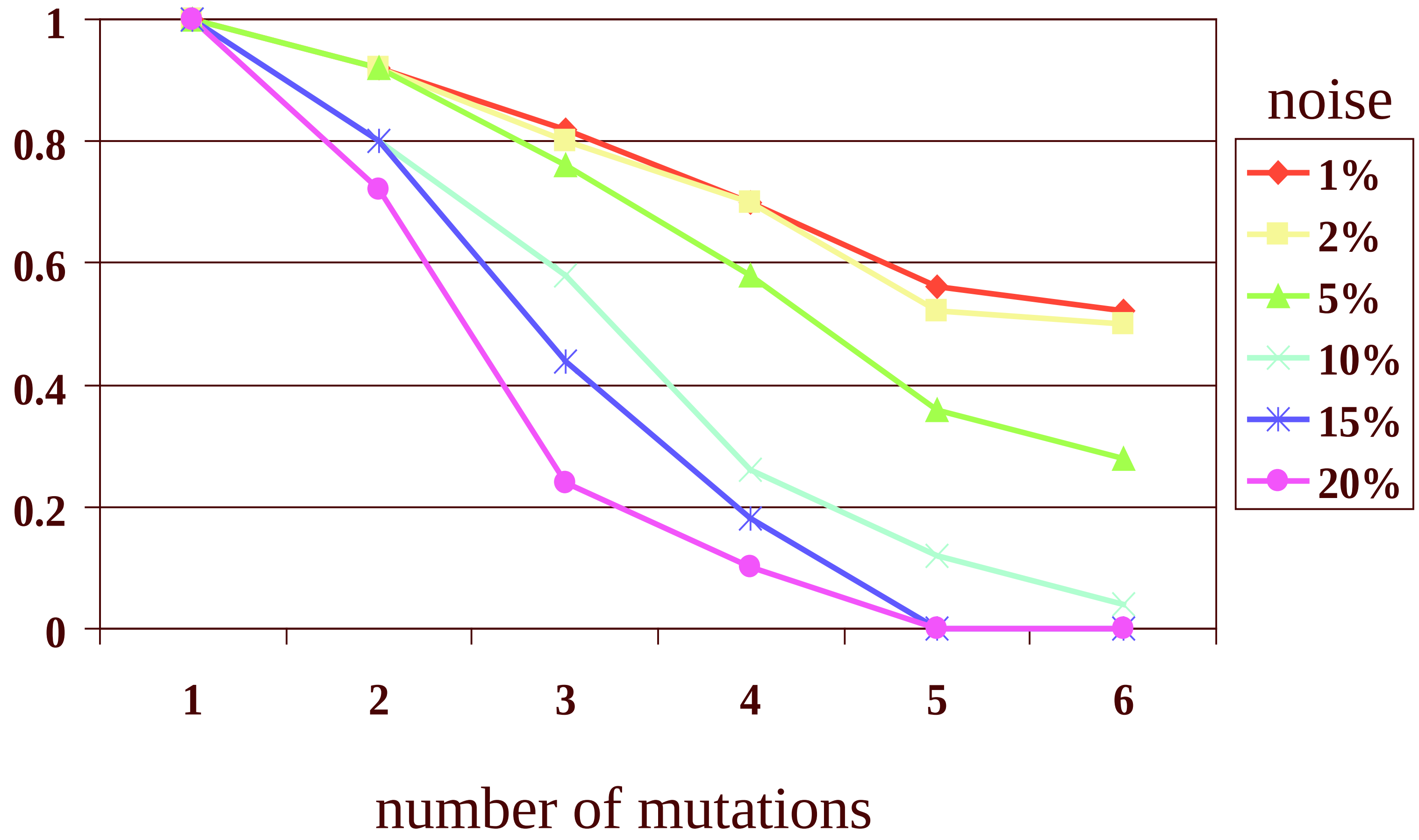
Simulated results

- p53 gene, exon 4
  - 167 substitutions (single & multiple)
  - 22 insertion
  - 76 deletion
- Mixes of up to 6 mutations + wildtype
- 1%-30% error
- Weights of [error/2, 0.6*numMut]

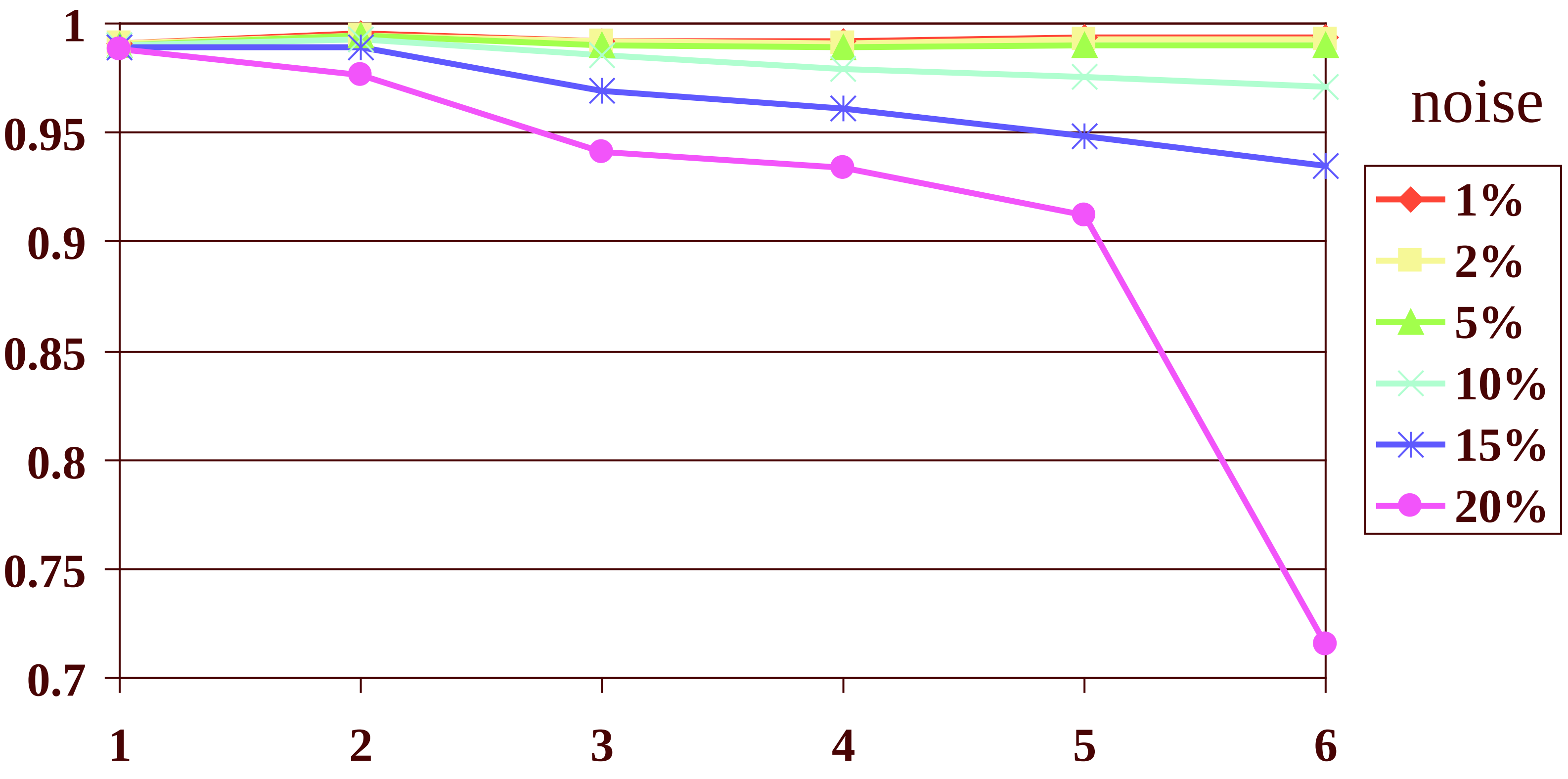
Likelihood at least 1 mutation correctly detected



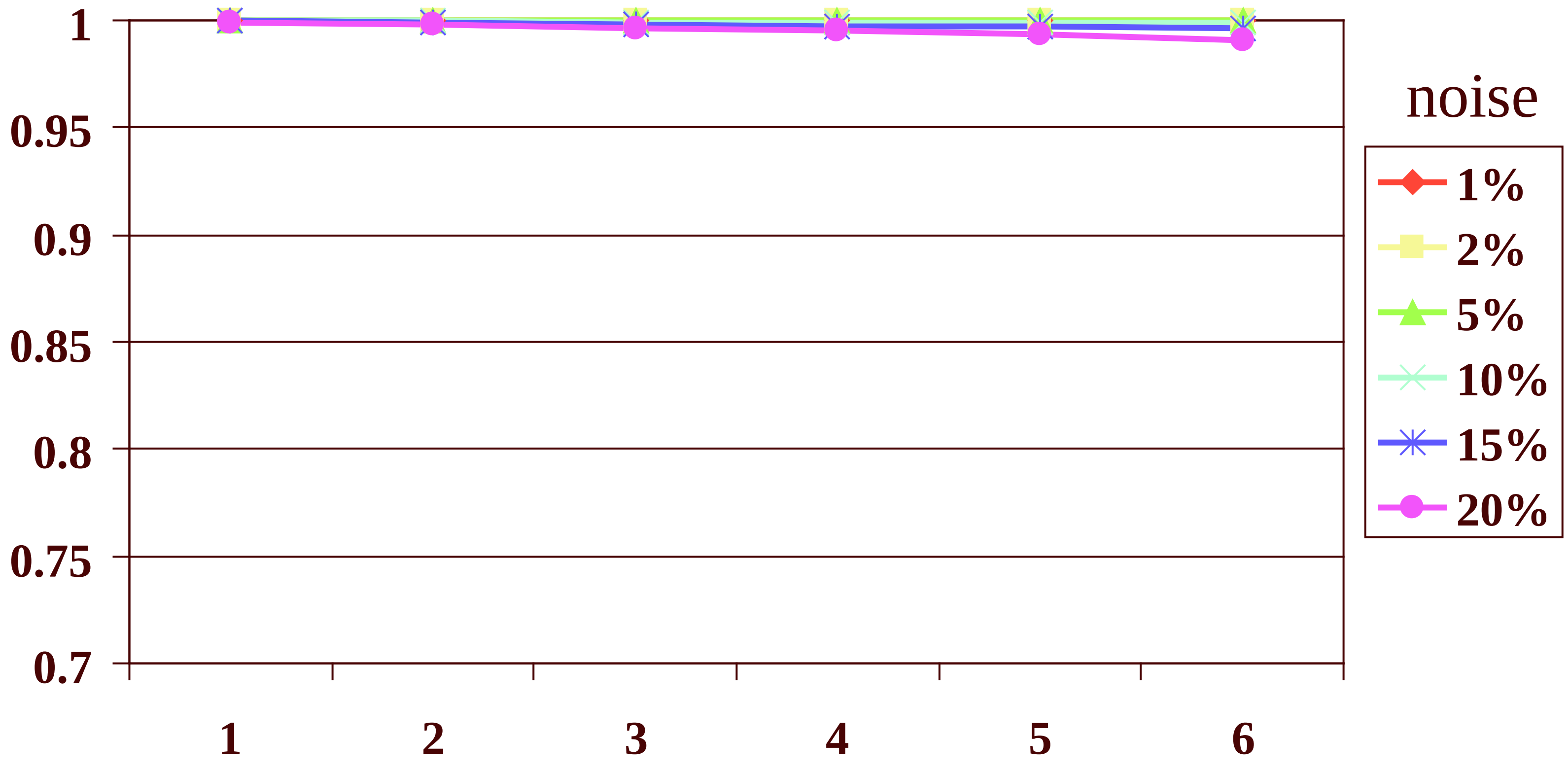
Likelihood all mutations correctly detected



Frequency correlation



Frequency correlation given correct deconvolution

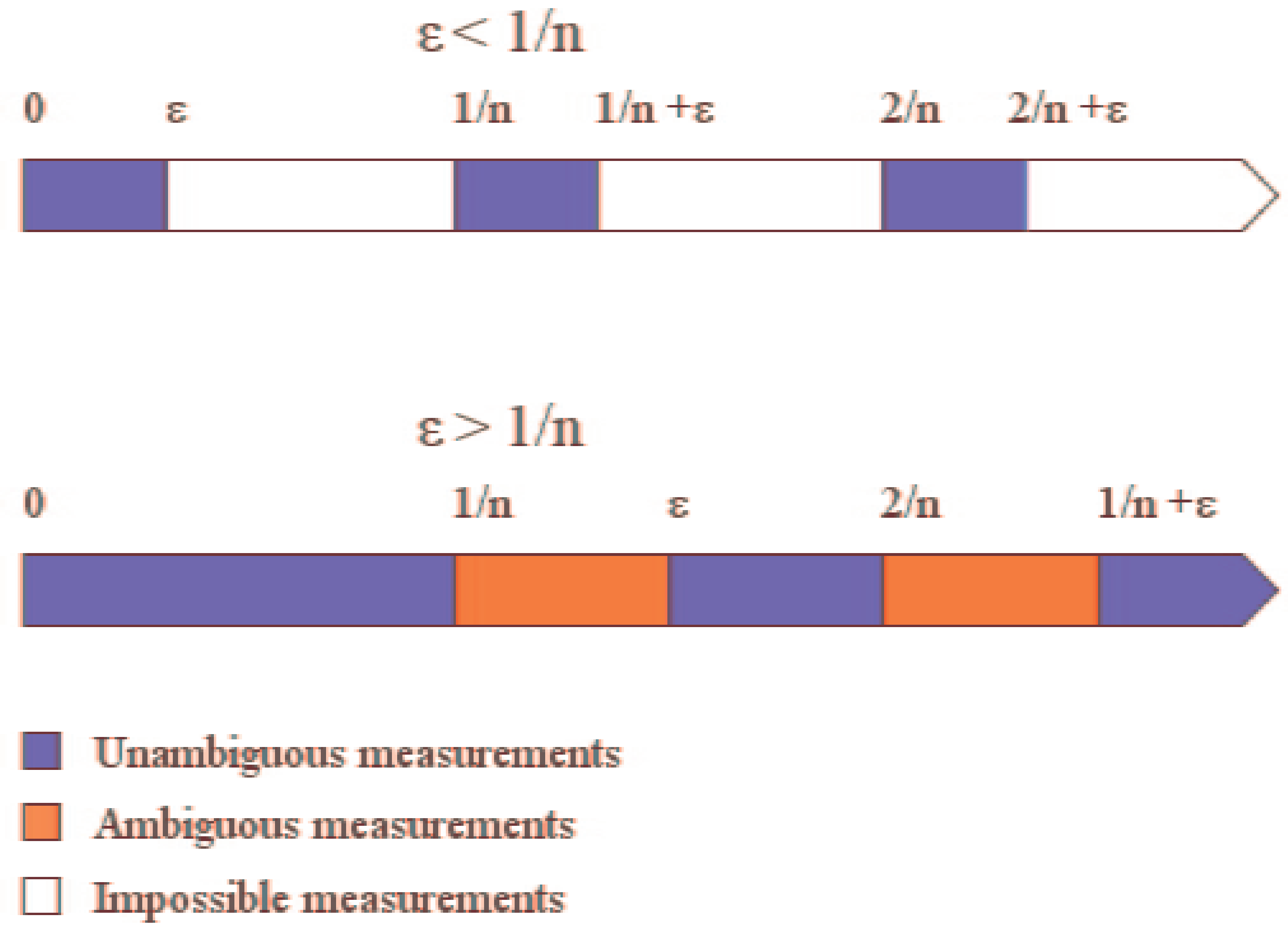


almost all error is from mistakes in deconvolution

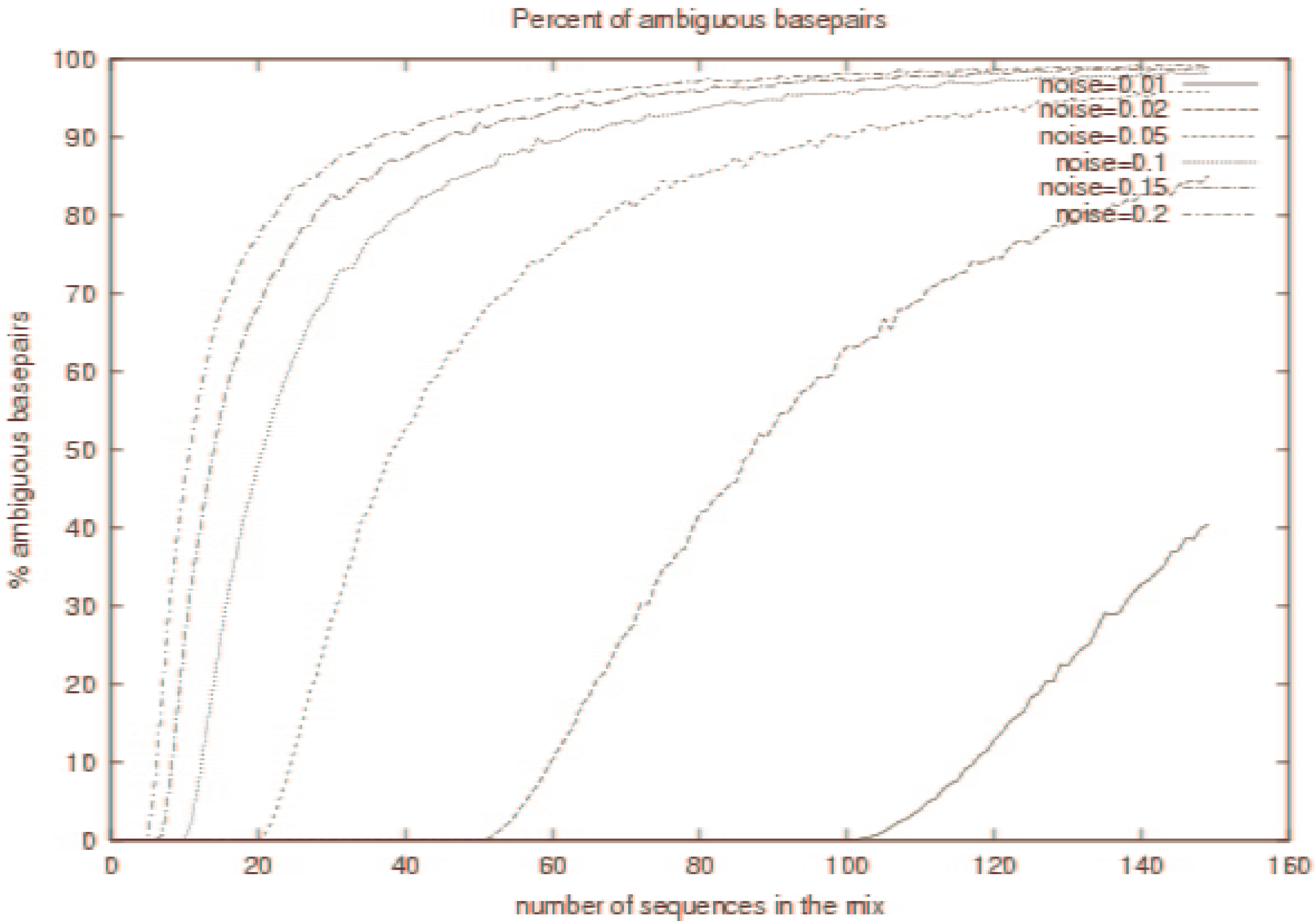
Detecting SNPs

- Detecting substitution mutations
- All mutations allowable
- O(n) trivial algorithm to detect them
- Increase throughput by mixing samples

Detecting SNPs



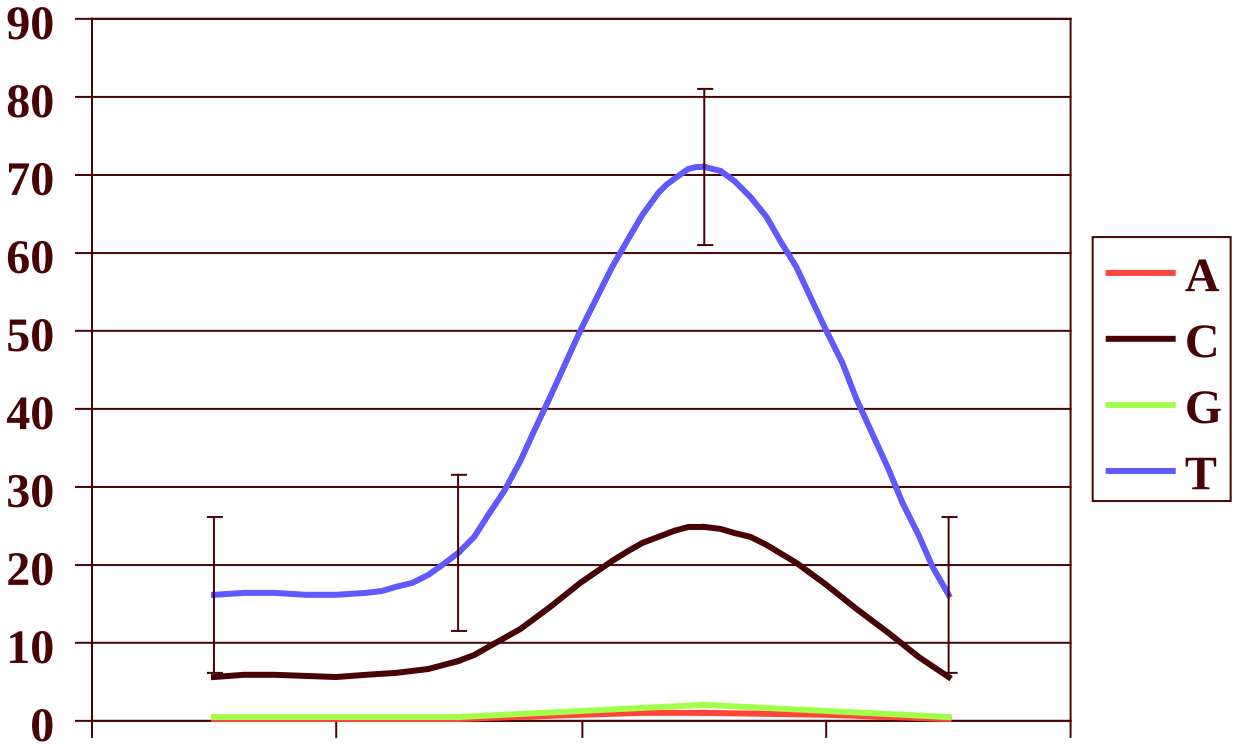
SNP Results


Percent of ambiguous basepairs
% ambiguous basepairs
number of sequences in the mix
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100
120
140
160
noise=0.01
noise=0.02
noise=0.05
noise=0.1
noise=0.15
noise=0.2

Future Directions

- Real experiments
- Improved noise models
- Different energy models
- Prior information on mutations

Questions


90
80
70
60
50
40
30
20
10
0
A
C
G
T

The future of Sanger sequencing

- Cheaper machines
- Longer sequences
- More accurate estimates at each basepair

Much cheaper sequencing

- physically small (8” x 8” x 16”)
- relatively cheap ($10k?)
- sequencer in every doctor’s office
- replace/supplement traditional lab tests

Idealized (future) Sanger sequencing

- Presence/absence of each base

```
ACTGTTGACTCATCCC
 AGTC  CTCATCG
```

- weight of each base at each position

```
Basepair 10: A=1% C=25% G=2% T=71%
```

Motivation

- Acquired mutations in cancer/virus
- sequencers in doctors office

Mutation Convolution

Sequence input

TGTTGACTCATCCC	Wildtype
TGTTCACTCATCCC	Sub(5,C)
TGAAGACTCATCCC	Sub(3,AA)
TGTTGACTCCCC	Del(10,2)
TGTTGCACTCATCCC	Ins(6,C)

Sequence output

```
TGTTGACTCATCCC     Wildtype
  AACCACTCCT  C    other
          A
```

Deconvolution can have many

profile

```
TGTTGACTCATCCC     Wildtype
  AACCACTCCT    C  other
        A
```

solution 1

```
     CACTCATCCC    Ins(6,C)
  AAC              Sub(3,AAC)
          C        Sub(11,C)
```

solution 2

```
     CACTCATCCC  Ins(6,C)
  AACCACTCC      Sub(3,AACCACTCC)
```

Sequencing Mixed DNA

- Base calling
- Mutation deconvolution
- Population frequency determination

Gel electrophorisis

- produce curves registering amount of each base at each position
- for homogeneous samples, “largest” peak defines underlying sequence
- for inhomogeneous samples, relative weights may yield info on presence/frequency of mutations

Goals

- Simultaneously detect multiple p53 mutations
- High-throughput method for detecting SNPs
- Viral population analysis

Three ways to solve

- Pseudo-inverse
  - min. squared error, allows negative weights
  - 4s linear equations -- fast
- Linear Programming
  - min. absolute error, weights non-negative
  - 4s constraints, 8s dummy variables -- slow
- Quadratic Programming
  - min. squared error, weights non-negative
  - 4s constraints, 4s dummy variables -- slower

Inhomogeneous sample

- relative weights may yield info on presence/frequency of mutations