Parametric Inference for Recombination in HIV Genomes

Kevin Woods, Oberlin College
(joint with Niko Beerenwinkel and Colin Dewey)

Genomes from two strains of HIV:

`AAAAAA`

`CCCCCC`

Find new strain:

`AAACCC`

Genomes from two strains of HIV:

AAAAAA

CCCCCC

Find new strain:

AAACCC

Recombination!

Goal: Given genomes of a number of parent strains and a new genome, figure out exactly how it is a recombination.

`AAA`

`CCC`

New strain:

`TAC`

Goal: Given genomes of a number of parent strains and a new genome, figure out exactly how it is a recombination.

`AAA`

`CCC`

New strain:

`TAC`

Is it

`TAC` ?

1 mutation, 1 recombination event

Goal: Given genomes of a number of parent strains and a new genome, figure out exactly how it is a recombination.

`AAA`

`CCC`

New strain:

`TAC`

Is it

`TAC` ?

1 mutation, 1 recombination event

Is it

`TAC` ?

2 mutations, 0 recombination events

Tradeoff between mutations and recombinations.

Two parameters R and M, the costs of a recombination event or a mutation.

Annotation: a coloring of the new sequence as a recombination of the parent strains.

Given an annotation with r recombination events and m mutations,

$$\text{total cost } = R \cdot r + M \cdot m$$

Minimize total cost over all possible annotations.

Two parameters R and M, the costs of a recombination event or a mutation.

Annotation: a coloring of the new sequence as a recombination of the parent strains.

Given an annotation with r recombination events and m mutations,

$$\text{total cost } = R \cdot r + M \cdot m$$

Minimize total cost over all possible annotations.

But what are R and M?

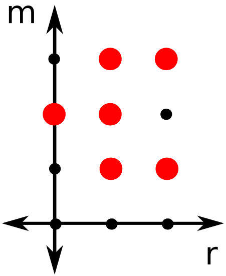
Parents:

`AAA`

`CCC`

New strain:

`TAC`



8 possible annotations.

Graph (r, m) for each annotation.

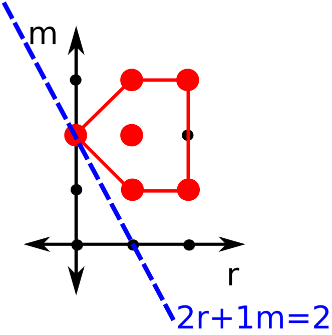
Example: $R = 2$, $M = 1$

Maximize

$$2r + 1m$$

over all points.

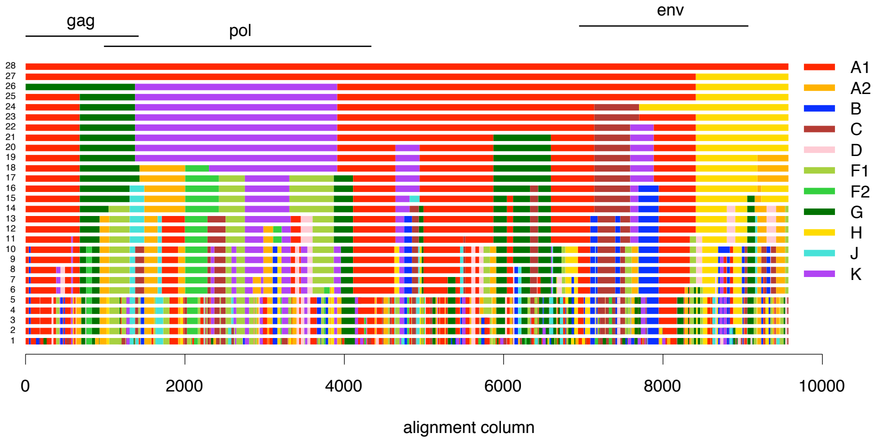
Linear Programming!

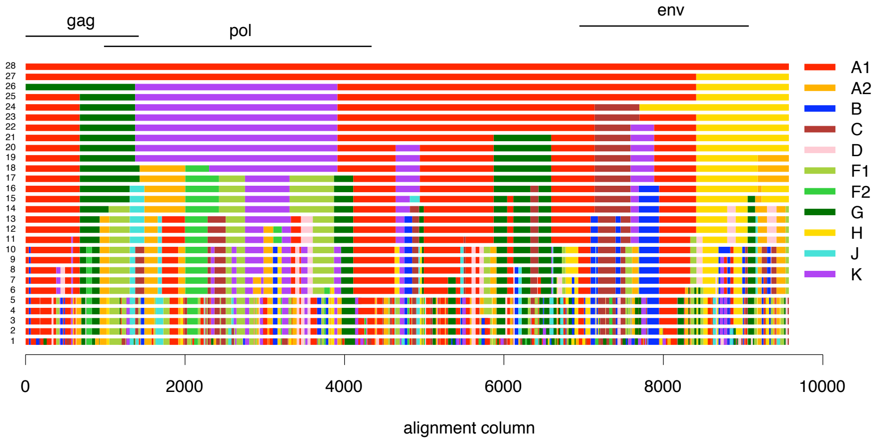


The vertices of this polytope are all of the information we need to understand this problem for all parameters (Pachter, Sturmfels).

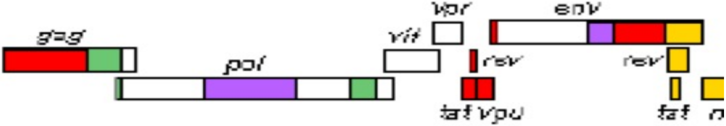
We can use all of our tools and algorithms from polytopal combinatorics.

This approach gives us a manageable number of interesting annotations.


gag
pol
env
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
A1
A2
B
C
D
F1
F2
G
H
J
K
0
2000
4000
6000
8000
10000
alignment column



gag
pol
env
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
A1
A2
B
C
D
F1
F2
G
H
J
K
0
2000
4000
6000
8000
10000
alignment column



vpr
env
gag
vif
pol
rev
rev
tat
vpu
tat
n

This can be translated to a probabilistic Hidden Markov Model, detailing

$$P(\text{annotation}, \text{child sequence} \,\Big|\, \text{parameters}, \text{parent sequences})$$

For fixed parameters “minimizing cost” is finding annotation that maximizes this probability (MAP estimate).

What are the parameters?

What are the parameters?

Classical answer: Maximum Likelihood Estimate: Find the parameters that maximize

$$P(\text{child sequence} \;\Big|\; \text{parameters}, \text{parent sequences})$$

What are the parameters?

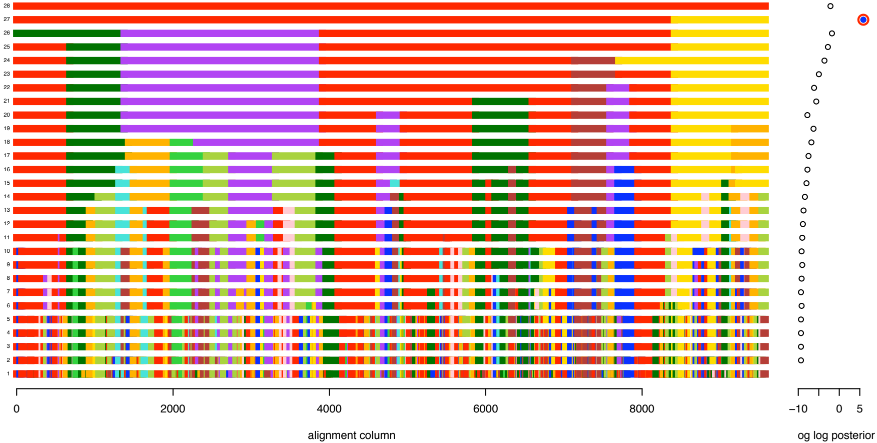
Classical answer: Maximum Likelihood Estimate: Find the parameters that maximize

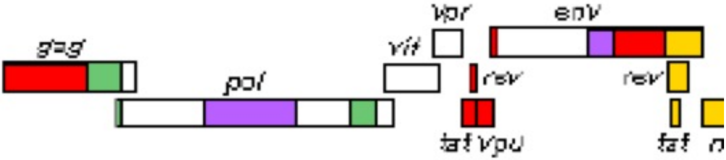
$$P(\text{child sequence} \mid \text{parameters, parent sequences})$$

Our approach allows interesting alternative: Given prior distribution on $P(\text{parameters})$ (uniform?), compute, for each vertex,

$$P(\text{vertex is MAP estimate} \mid \text{parent sequences, child sequence})$$

This approach gives us a manageable number of interesting annotations, each with a numerical score associated with them.


28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
2000
4000
6000
8000
alignment column
−10
0
5
og log posterior


gag
pol
vif
vpr
env
rev
tat
vpu
rev
tat
n