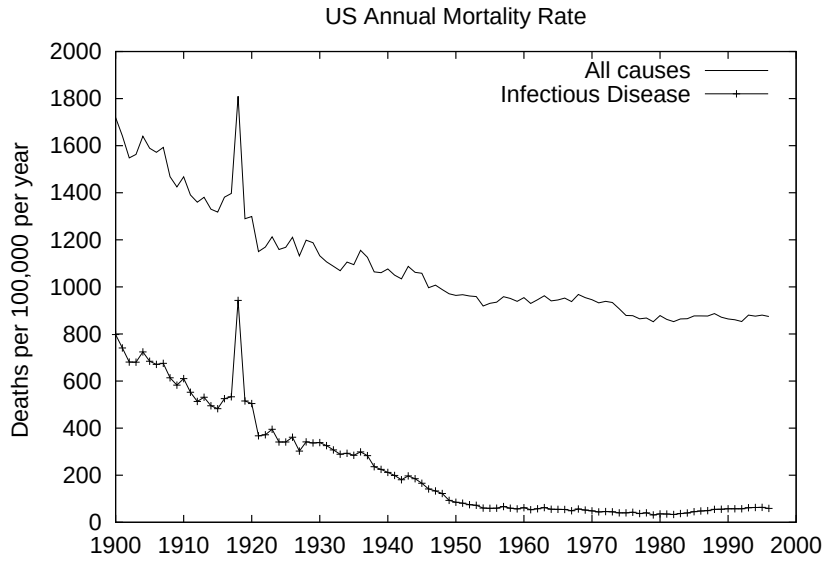
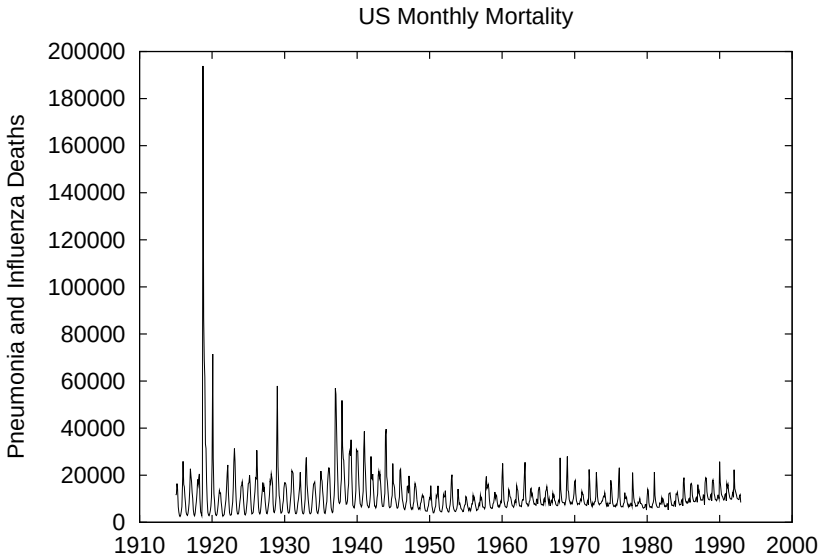
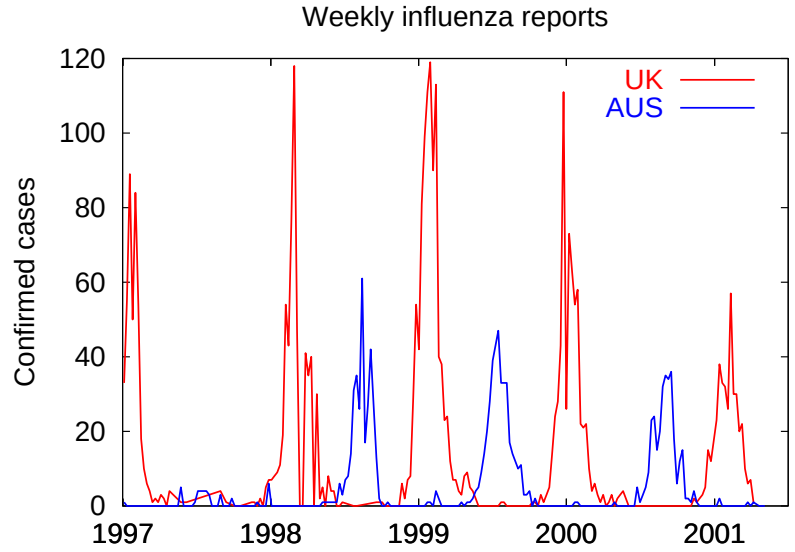
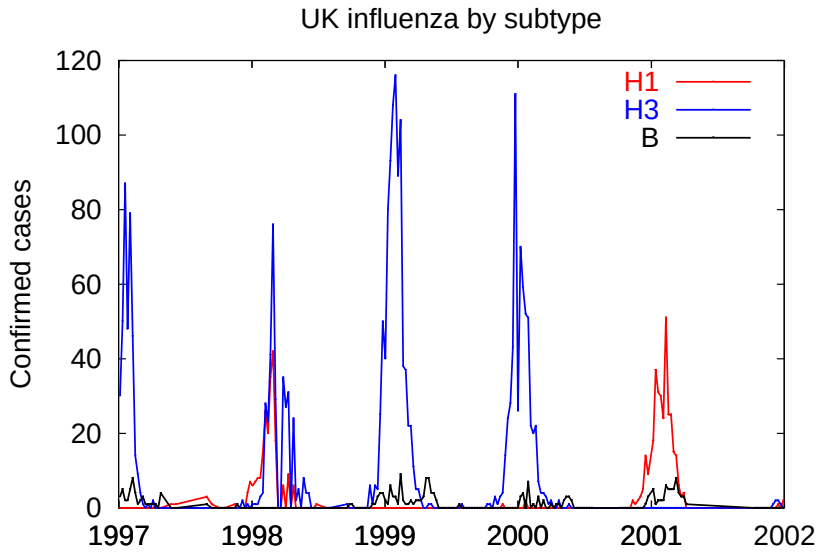
# Patterns of hemagglutinin evolution and the epidemiology of influenza

*DIMACS Working Group on Genetics and Evolution of Pathogens, 25 Nov 03*






## Influenza viruses

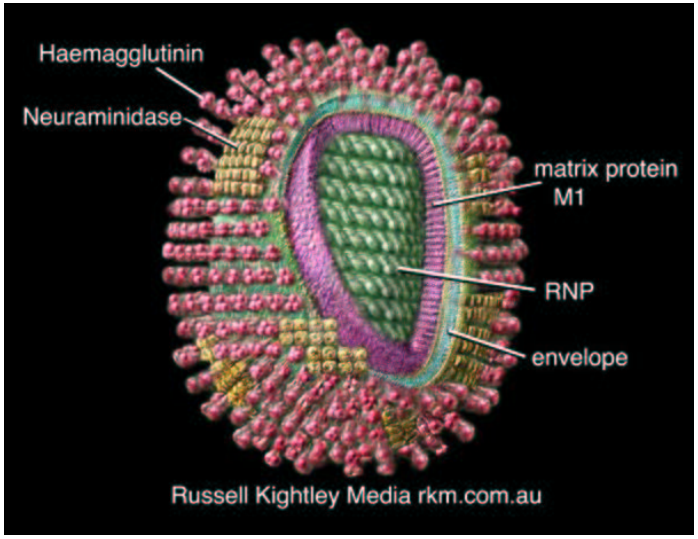
Three *types,* A, B and C, in decreasing order of importance.

Flu A has fifteen identified hemagglutinin *subtypes*, all of which are always present in waterfowl.

Evolutionary *shifts* occur when core proteins from human-adapted strains recombine with surface proteins from avian strains, probably in people, domestic fowl or pigs.

Evolutionary *drift* in the surface proteins means that most people are susceptible to a related, circulating strain of the flu around five years after recovery.

An influenza virion




## Shift evolution

Major antigenic change caused be reassortment between human and avian virus segments.

**1918** Spanish flu (H1N1) replaces earlier strain.

**1957** H2N2 replaces H1N1.

**1968** H3N2 replaces H2N2.

**1977** H1N1 mysteriously reappears.

It is estimated that there have been roughly 10 influenza pandemics (presumably caused by shifts) in the last 250 years.

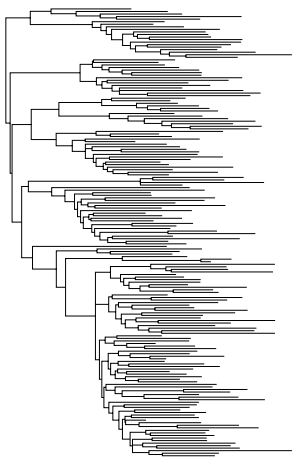
## Drift evolution

The gradual accumulation of point mutations antibody-combining regions (epitopes), leading to immunological escape.

Makes vaccine-strain selection very difficult.

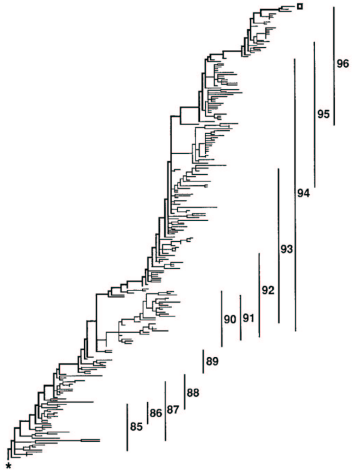
Annual epidemics due to drift cause more total mortality and morbidity than pandemics.

HIV-1



Rambaut, et al., 2001

Influenza A




Fitch, et al., 1997

## Why model infectious diseases?

How do local interactions explain population-level patterns?

What can population-level patterns tell us about local interactions?

## Questions to be addressed by influenza modelling

How do different subtypes interact at the population level, and what can this tell us about pandemics?

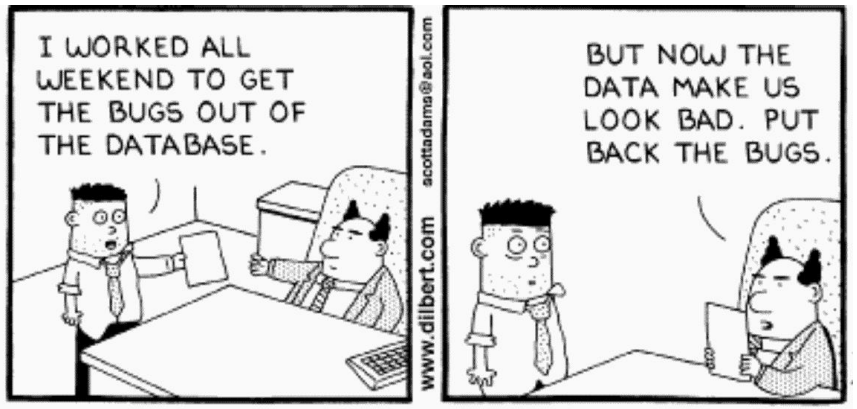
What factors determine influenza's unique phylogenetic patterns?

Can predictions about drift evolution improve annual vaccine choices?

Why does influenza incidence show such marked seasonal oscillations?

What are the implications of influenza's antigenic evolution for drug resistance?

Confronting models with data




## Outline

Clustering

More clustering

Volatility

More volatility

## Quasispecies structure and the antigenic evolution of Influenza A

*Joshua Plotkin, Jonathan Dushoff, Simon Levin; PNAS 99:6263*

**Questions**

- What do modelers mean by a 'strain'?
- What does strain space look like?
- Do influenza viruses cluster into 'quasispecies'?

## How to compare hemagglutinin molecules

Antigenic assays

Three-dimensional structure

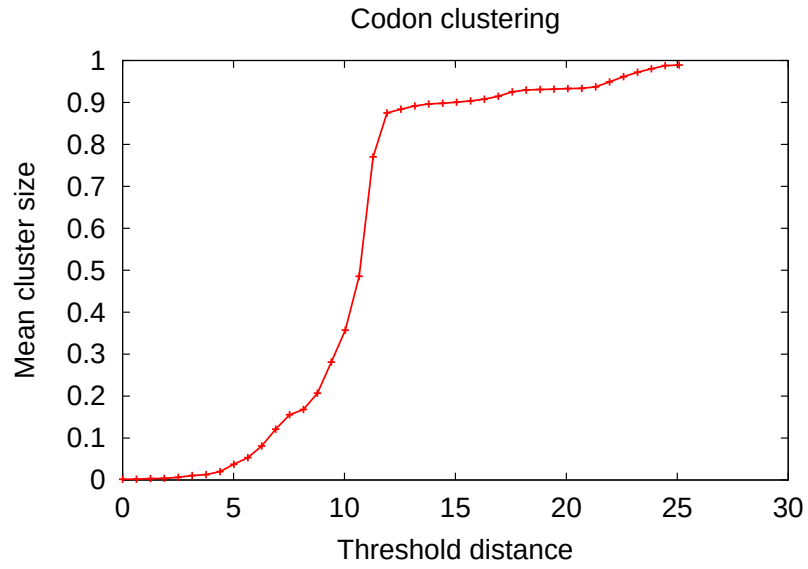
*Amino-acid sequence*

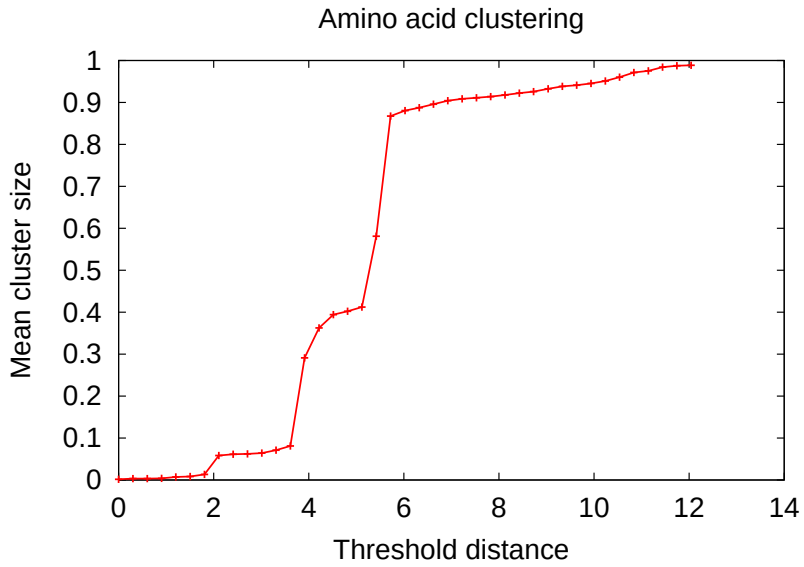
- Simple
- Precise
- Available

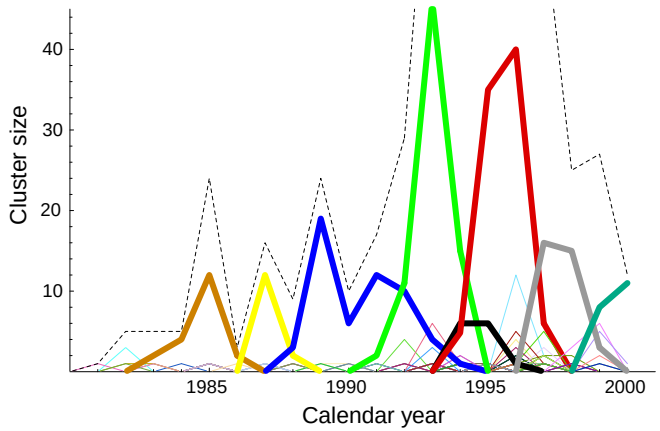
## Random clustering technique

Examine random clusters at different length scales

Look for scales at which clusterings are stable; these are natural clusterings

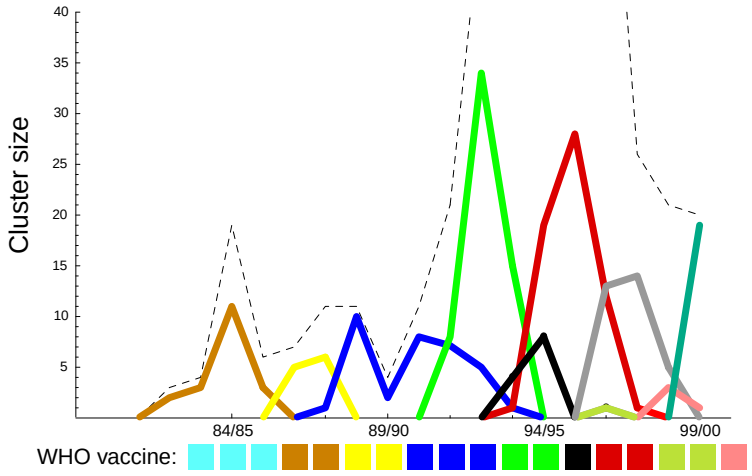


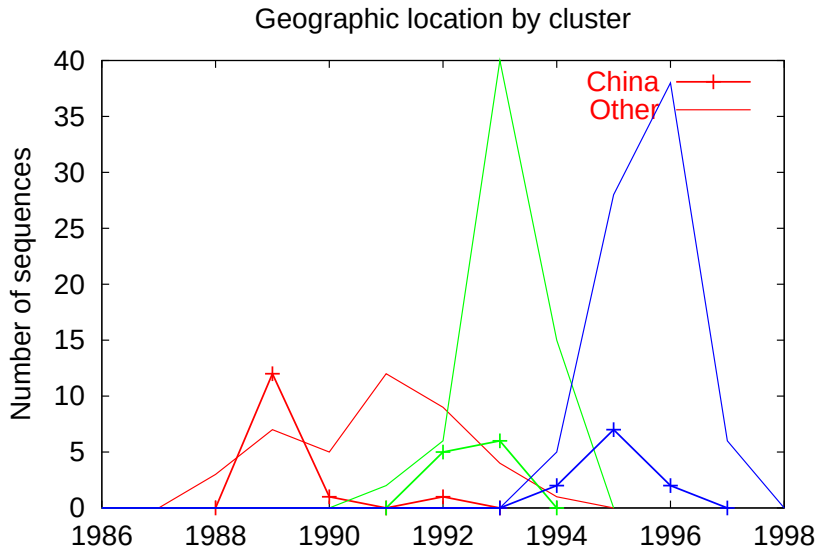



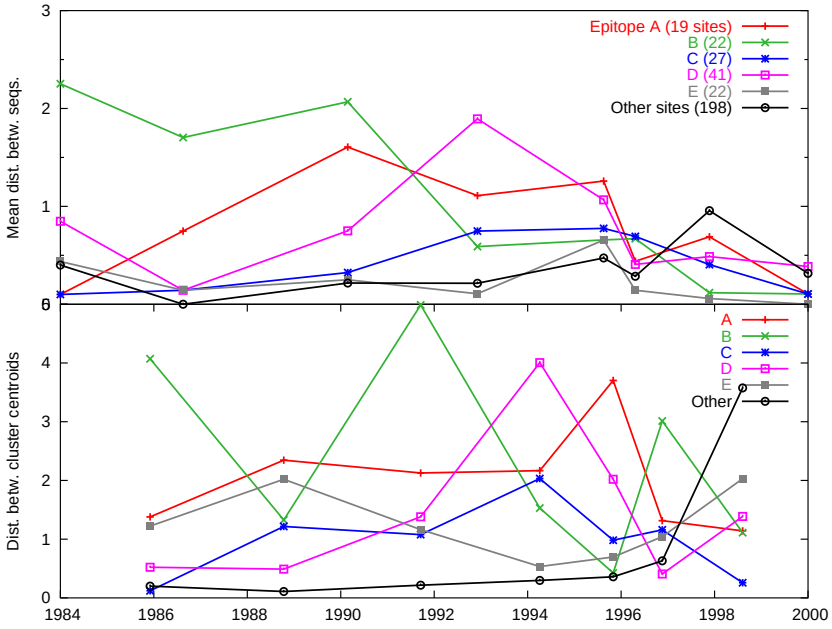



## Clusters through time

- Quasispecies have limited temporal range
- Dominant quasispecies replace each other on a time scale of 2–5 years
- Evolution is linear over this time span in amino-acid space



## Conclusions

- Sequences are clustered in amino-acid space, forming natural 'quasispecies'.
- Clusters replace each other on a time scale of 2–5 years.
- Clusters display interesting interactions with antibody-combining regions (epitopes).
- Formal clustering methods have potential for predicting the direction of influenza evolution.

## Homology modeling

Start with backbone in the same place as known structure.

Adjust for stereochemical constraints and known motifs.

Local energy minimization.

*Works surprisingly well over a broad range of proteins.*

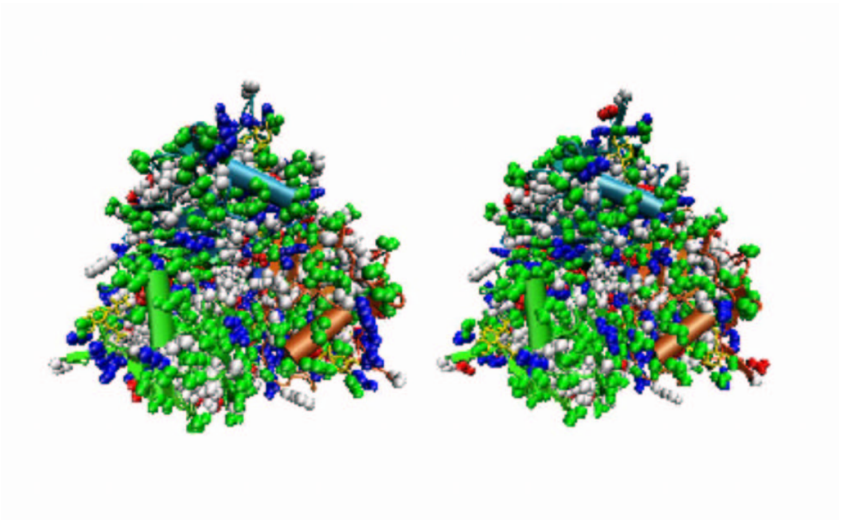
## Clustering methods for HA structures

*with Ben McMahon and Joshua Plotkin*

If formal clustering methods assist in analysis of genomic patterns, can they also assist in analysis of structural patterns?

Computational and algorithmic advances make it possible to make homology models for hundreds of sequences (based on known structures).

Human H3 structures



The cartoon shape is the backbone (with a color gradient).

Spheres are the side chains:

**White** Non-polar

**Blue** Positive

**Red** Negative

**Green** Other polar

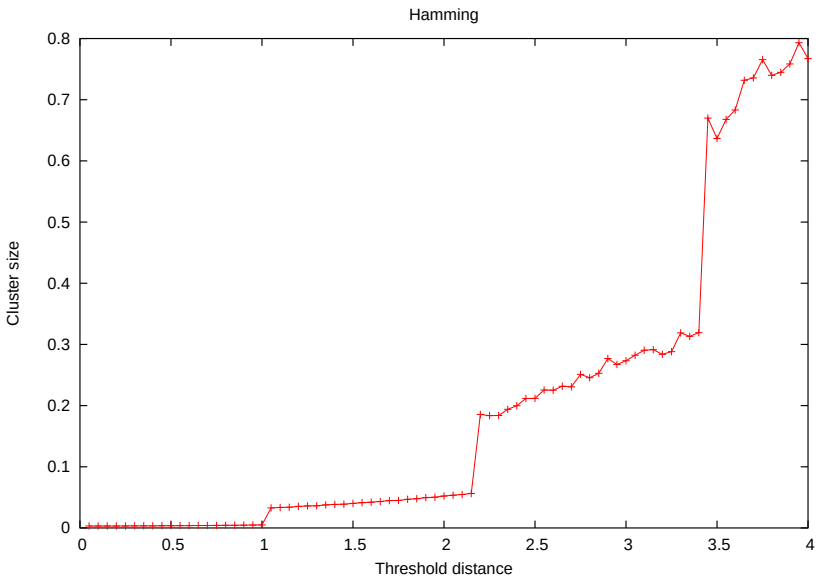
The yellow is a sialic acid analog bound to the protein.

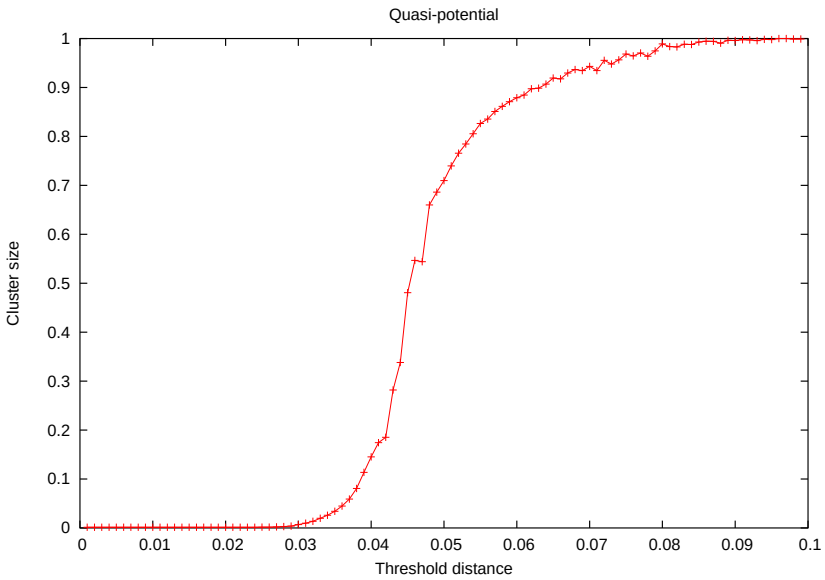
## Clustering methods for HA structures

*Relative methods*

Calculate profiles based on the protein backbone (e.g. the electric field at each of the 329 alpha carbons).

- Hydrophobicity
- Electric field and potential measures
- Distance profiles

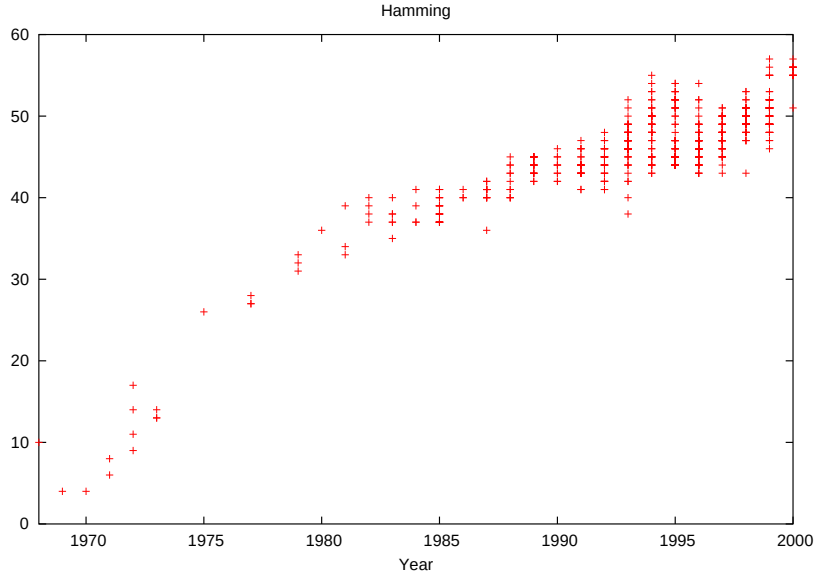


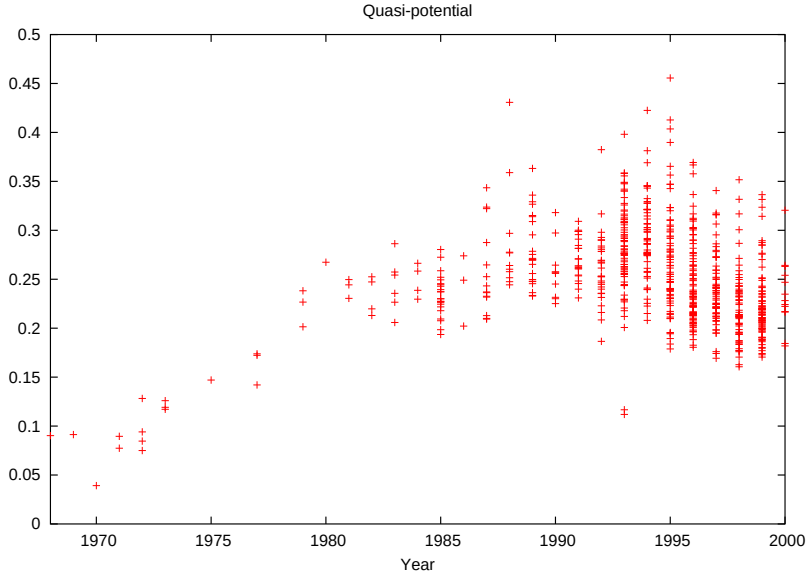



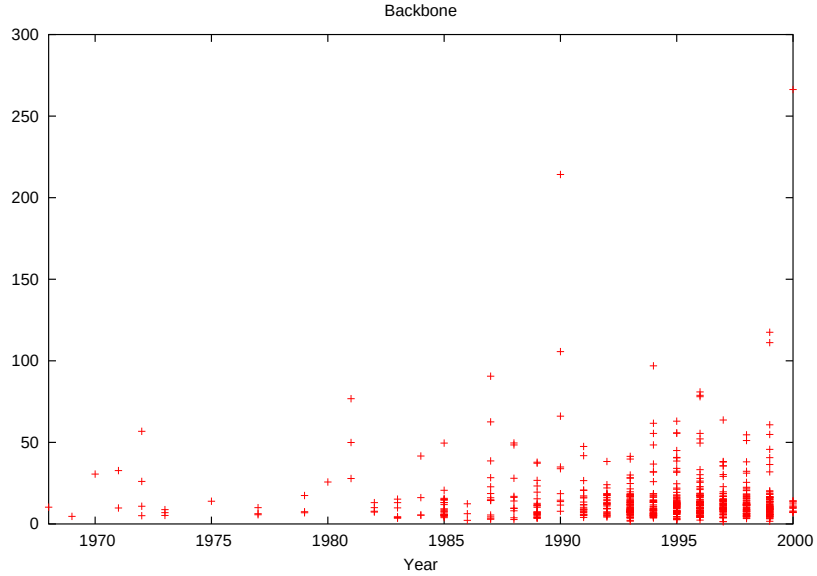
## No natural scale for structural clusters

Need more sophisticated clustering methods

Evidence for compensatory mutations



## Provisional conclusions

Evidence for existence of compensatory mutations.

Simple, relative measures, combined with homology models, may be able to detect and explain these compensatory mutations.

More refined metrics needed to cluster in ways that will shed light on antigenicity.

More refined clustering techniques may also be needed.

## Natural selection in pathogens

**Stabilizing selection** (selection not to change) implies inflexibility, importance.

**Positive selection** (selection to change) implies pressure from host immune system, or directional change (change of disease mechanism, or change of host)

Useful for investigating biology and evolution of pathogens

Potential applications for vaccine and drug development

## Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus

*Joshua Plotkin and Jonathan Dushoff; PNAS 100:7152*

**Questions**

- Can codon usage help to explain how hemagglutinin evolves so quickly?
- Does hemagglutinin's fast evolution leave a 'footprint' on codon usage?
- Can we correlate genomic information about evolution with structural information about hemagglutinin?

## Codon bias

Genomes use certain codons in high proportions, in preference to other, synonymous codons. This is surprising because there is no obvious reason why the organism should distinguish between synonymous codons.

Some reasons for codon bias include:

- Nucleotide biases
- Mutational biases
- The mechanics of translation
- *Evolutionary history*

## Bias towards volatility

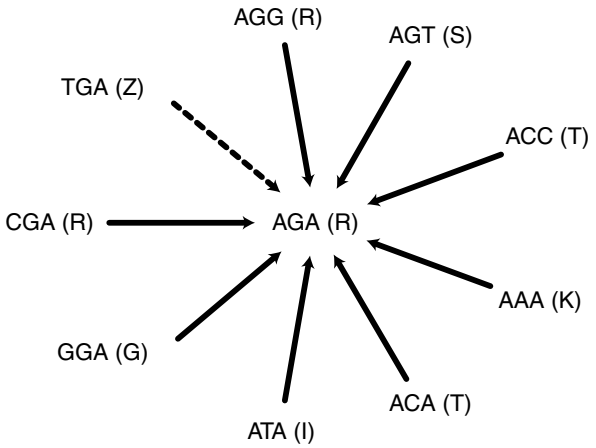
Some codons have more synonymous neighbors than others. Under *neutral* selection, all of the non-stop neighbors of a codon are equally likely as predecessors.

If a gene is under *positive* selection, the predecessor codon is more likely to have been non-synonymous.

If a gene is under *negative* selection, the predecessor codon is more likely to have been non-synonymous.

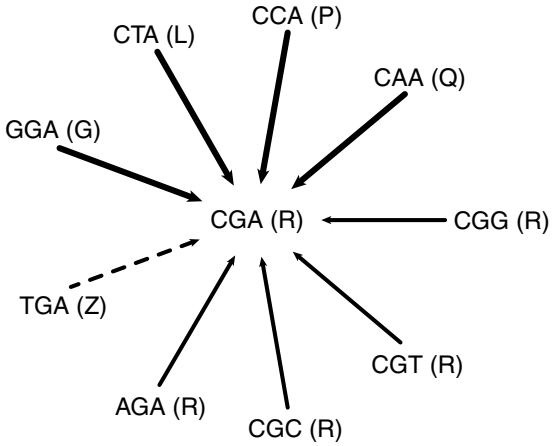
Thus, an overabundance of codons with more non-synonymous neighbors (high *volatility*) is a marker of positive selection. And conversely.

## Volatility




8 non-stop neighbors. 6 encode other amino acids (non-synonymous changes). Volatility = 6/8.

## Volatility




8 non-stop neighbors. 4 encode other amino acids (non-synonymous changes). Volatility = 4/8.

## Detecting bias towards volatility

Problem: Other sources of codon bias. Amino-acid composition of genes will bias measures of volatility.

Solution: Control for amino-acid composition by making bootstrap copies of the gene, with the same amino-acid composition.

```
ATG GAG AGC CTT GTT CTT GGT GTC AAC GAG AAA ACA  (Gene)

 M   E   S   L   V   L   G   V   N   E   K   T   (Protein)

ATG GAG AGC CTT GTT cta ggc GTC aat GAG AAA act  (Copy)
```

A gene shows significantly high (or low) volatility if its volatility exceeds (is below) that of 97.5% of the bootstrap copies.

## Volatility results

- Surface protein hemagglutin significantly volatile compared to rest of genome
- Antibody-binding areas of hemagglutinin significantly volatile compared to rest of hemagglutin
- No evidence that neuraminidase (the other surface protein) is volatile.

*Confirms that antibody-binding areas of hemagglutinin are under pressure to evolve continually, due to selective pressure from the immune system.*

## Bootstrap method

Controls for:

- nucleotide bias
- codon-specific bias
- Differences in underlying mutation rate

Does not control for:

- Expression bias
- Biases localized to certain parts of the genome

Is weakened by:

- Transition-transversion bias
- Specific highly mutable 'motifs'

*Currently more appropriate for pathogens than people*

## Comparing different amino acid metrics

**Hamming** Acids are the same, or different

**Miyata** Measures differences in size and hydrophobicity.

Comparisons of epitopes to non-epitopes are much less significant under the Miyata metric, likely reflecting structural constraints.

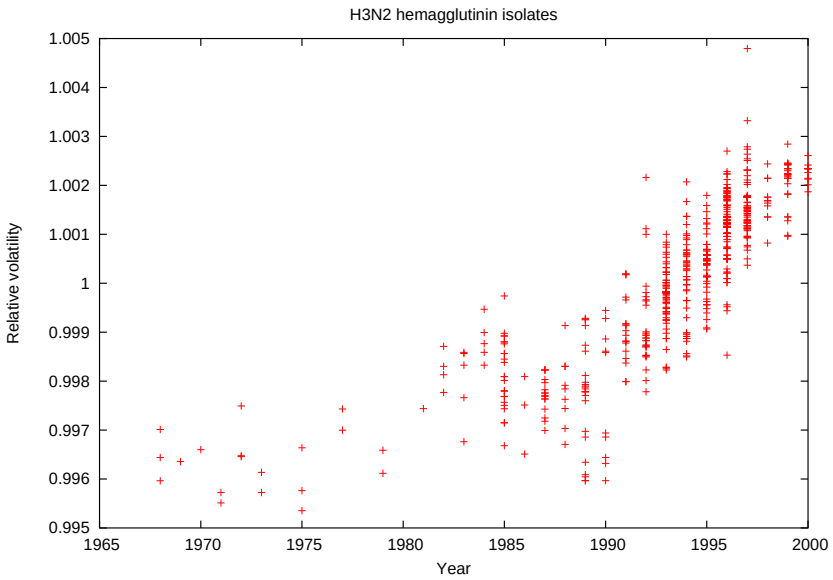
## Comparative hemagglutinin volatility

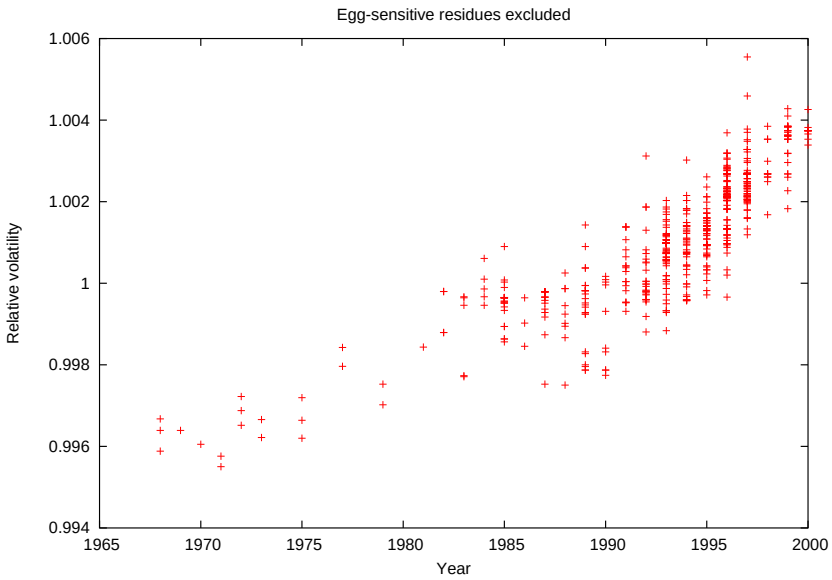
*with Joshua Plotkin*

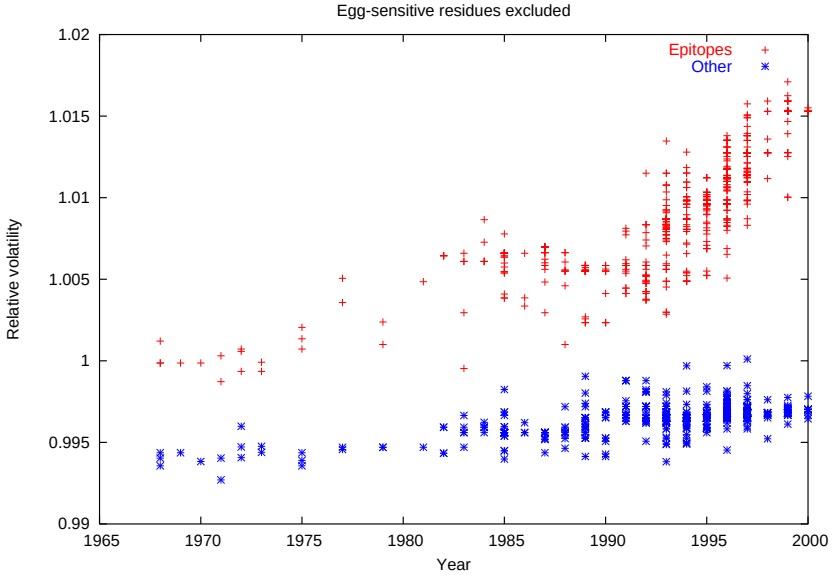
## Questions

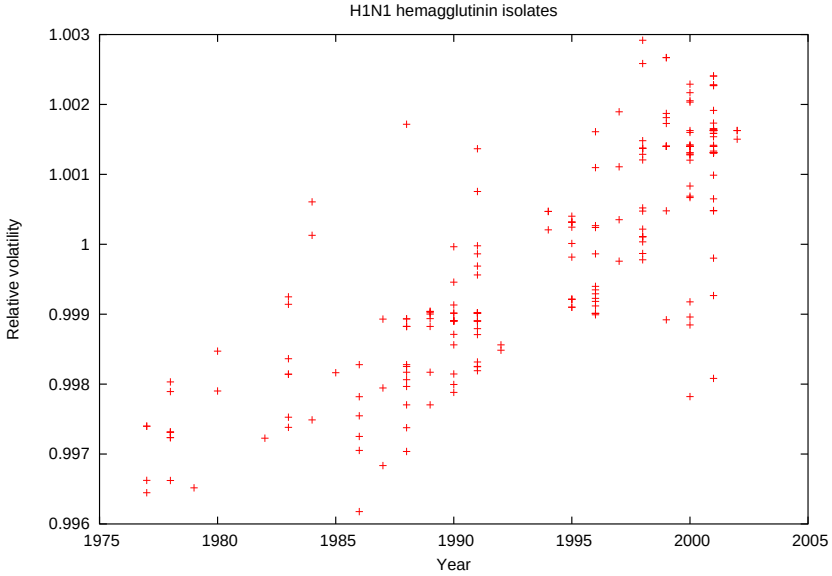
Is volatility really sensitive enough to distinguish between different phenotypes of the same gene?

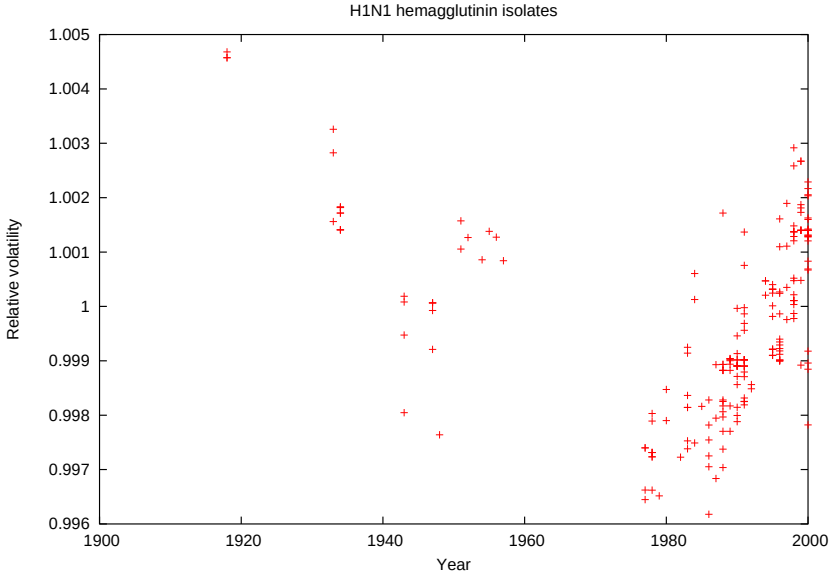
Can we learn anything about shifts from volatility patterns in pandemic isolates?

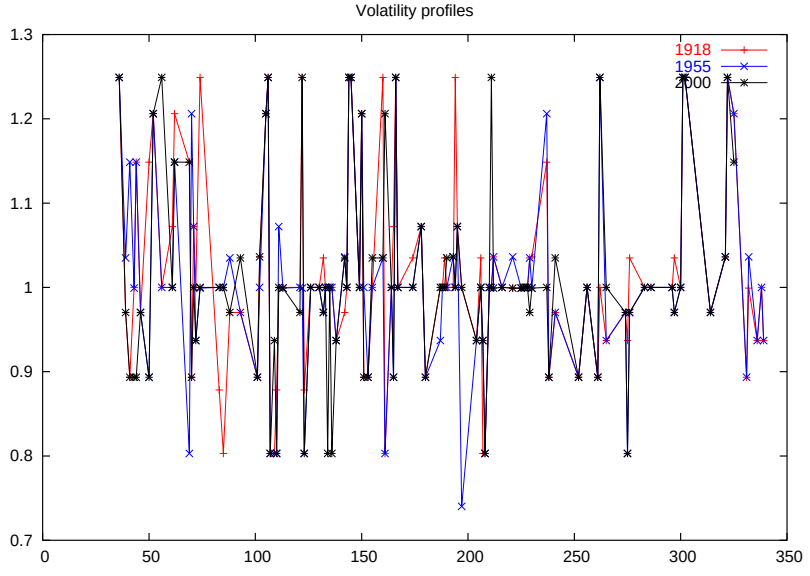



Egg-sensitive residues excluded
Epitopes
Other
Relative volatility
Year
1.02
1.015
1.01
1.005
1
0.995
0.99
1965
1970
1975
1980
1985
1990
1995
2000


H1N1 hemagglutinin isolates
Relative volatility
Year
1.003
1.002
1.001
1
0.999
0.998
0.997
0.996
1975
1980
1985
1990
1995
2000
2005


H1N1 hemagglutinin isolates
Relative volatility
Year
1.005
1.004
1.003
1.002
1.001
1
0.999
0.998
0.997
0.996
1900
1920
1940
1960
1980
2000


Volatility profiles
1918
1955
2000
1.3
1.2
1.1
1
0.9
0.8
0.7
0
50
100
150
200
250
300
350

## Conclusions

Not yet known if volatility is a sharp enough tool for this task.

Stay tuned