Bari, SM&FT 2011

Partitions and Rohlin's Distance: general formalism and applications to the evolution of Influenza A virus.

Raffaella Burioni (Dip. Fisica, Università di Parma)
Mario Casartelli (Dip. Fisica, Università di Parma)
Riccardo Scalco (Dep. Biochemistry, Zurich)

Why Partitions? Why Rohlin's Distance?

For any set **X**, a finite partition

$$\alpha = (A_1, A_2, A_3, \ldots, A_n)$$

is a covering of **X** by disjoint subsets (the "atoms" of α).

The interesting case:

When $\mathbf{X} = (\mathbf{M}, \mathscr{M}, \boldsymbol{\mu})$, i.e. a probability space,

and the atoms A_k are in $\mathscr{M}$, a partition represents

a probabilistic experiment with atoms $\{A_k\}$, k=1,2,…,n,

as outcomes having probabilities $\boldsymbol{\mu}(A_k)$.

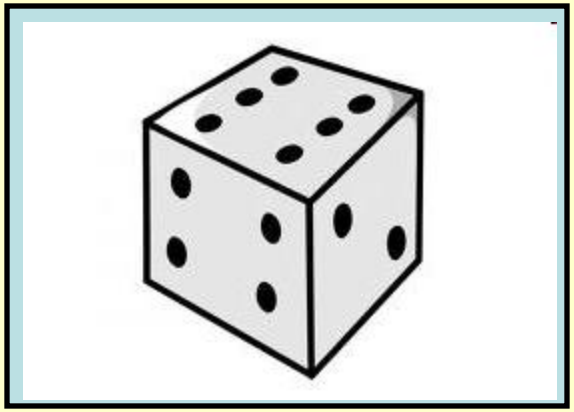
For instance, if **X** = (1,2,3,4,5,6) are the points-events of a die, the partition

α= ({1,3,5}, {2,4,6})

represents the **odd-even** experiment,
the partition

β = ({1}, {2,3,4,5,6})

the **“1” or “not-1”** experiment, etc.
The special partition v = (**X)** is the unit experiment (something always happens, information 0).
In finite spaces, also the Ɛ partion into single points may be considered. Here, Ɛ = {(1),(2),(3),(4),(5),(6)}



Partitions and Entropy

The Shannon's Entropy of a partition:

$$H(\alpha) = -\sum_{i=1}^{m} \mu(A_i) \ln \mu(A_i) \ .$$

It is the **mean information** necessary to know the results of an experiment

The conditional Shannon's Entropy

For any two partitions α and β, the conditional entropy H(α|β) is the residual incertitude about α when the result of β is known

$$H(\alpha|\beta) = -\sum_{i=1}^{m}\sum_{k=1}^{s}\mu(A_i \cap B_k)\ln\frac{\mu(A_i \cap B_k)}{\mu(B_k)} .$$

Rohlin's Distance in $\mathcal{Z}$

The "Partition Space" $\mathcal{Z}$ is the set of all finite measurable partitions. The Rohlin's distance $\mathbf{d_R}$ is a metrics in $\mathcal{Z}$ given for any couple of partitions α and β by the simmetrized conditional entropy

$$\mathbf{d_R(\alpha,\beta) = H(\alpha|\beta) + H(\beta|\alpha)}$$

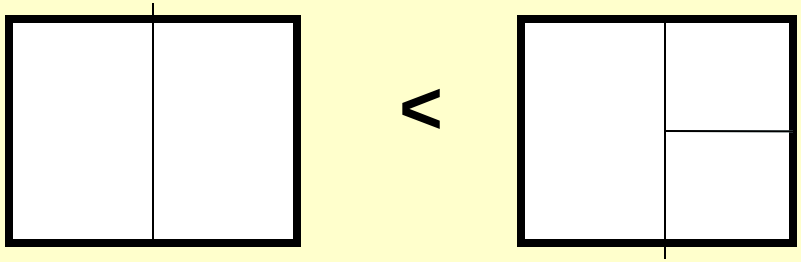
It is an index of **"non-similarity"**

Algebra on partitions

In order to compute distances, it is useful or necessary to exploit some algebraic features of $\mathcal{Z}$

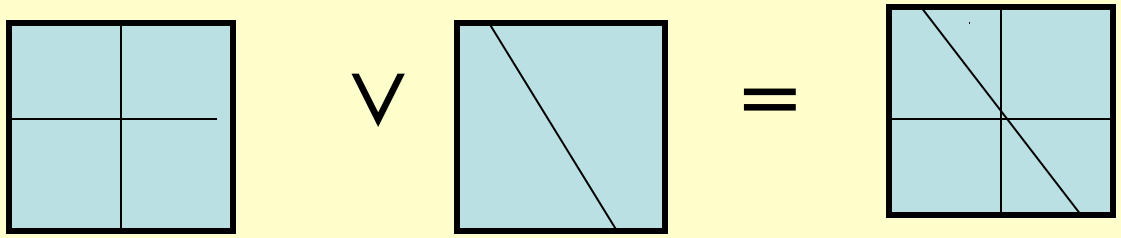
First, there is a partial order :

α < β means that **β** refines **α**.



Composition, or pseudo-product

γ = α∨β (or simply **γ = αβ**) is the minimal partition refining both α and β



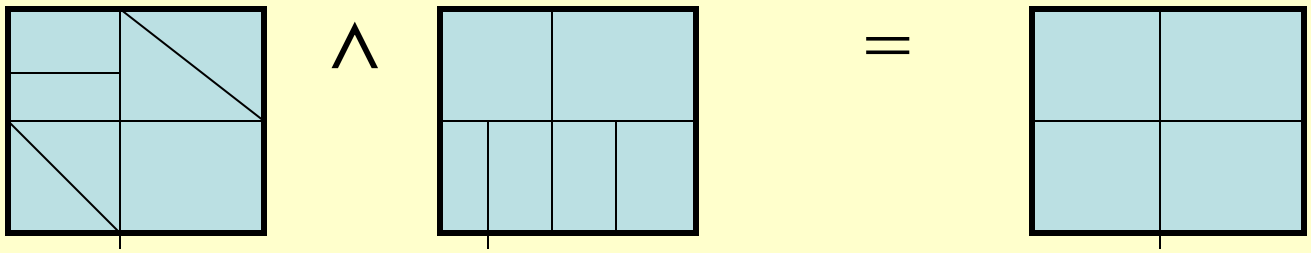
It is a "**minimal common multiple**" :

(Its atoms are non empty intersections of the factors atoms)

Intersection

σ = α∧β is the maximal sub-partition of both **α** and **β**, i.e. their

"maximal common factor"



An useful formula: $\mathbf{d_R(\alpha,\beta) = 2H(\alpha\beta) - H(\alpha) - H(\beta)}$

The Reduction Process

It is possible to amplify the non-similarity between partions by erasing as far as possible the common sub-partitions, or factors.

This process is not univocal, because the factorization into “prime factors” is not uniquely defined.

Dicothomic sub-partions may be considered as “prime” (i.e. indecomposable) factors, but they are extremely redundant:

2^{N-1} -1 for an *N*-atoms partition.

Elementary Factors

A "good" family $\mathcal{E}(\alpha)$ of dichotomic factors α_k is well defined if:

i – $\mathcal{E}(\alpha)$ may be defined for every α

ii – there are as many α_k as atoms in α

iii – $\mathbf{V}_1^N \alpha_k = \alpha$

The universal choice: $\alpha_k = (A_k, A_k^c)$

There are alternative choices in particular cases

Reduction

For every **α** and **β**, $\mathcal{E}(\boldsymbol{\alpha})$ and $\mathcal{E}(\boldsymbol{\beta})$ are defined.

Let $\boldsymbol{\sigma = \alpha \wedge \beta}$.

Recipe:

i - Drop from $\mathcal{E}(\boldsymbol{\alpha})$ the factors $\boldsymbol{\alpha_k}$ such that $\boldsymbol{\alpha_k \wedge \sigma \neq v}$.

ii - Let $\boldsymbol{\alpha'_k}$ be the surviving factors of **α.**

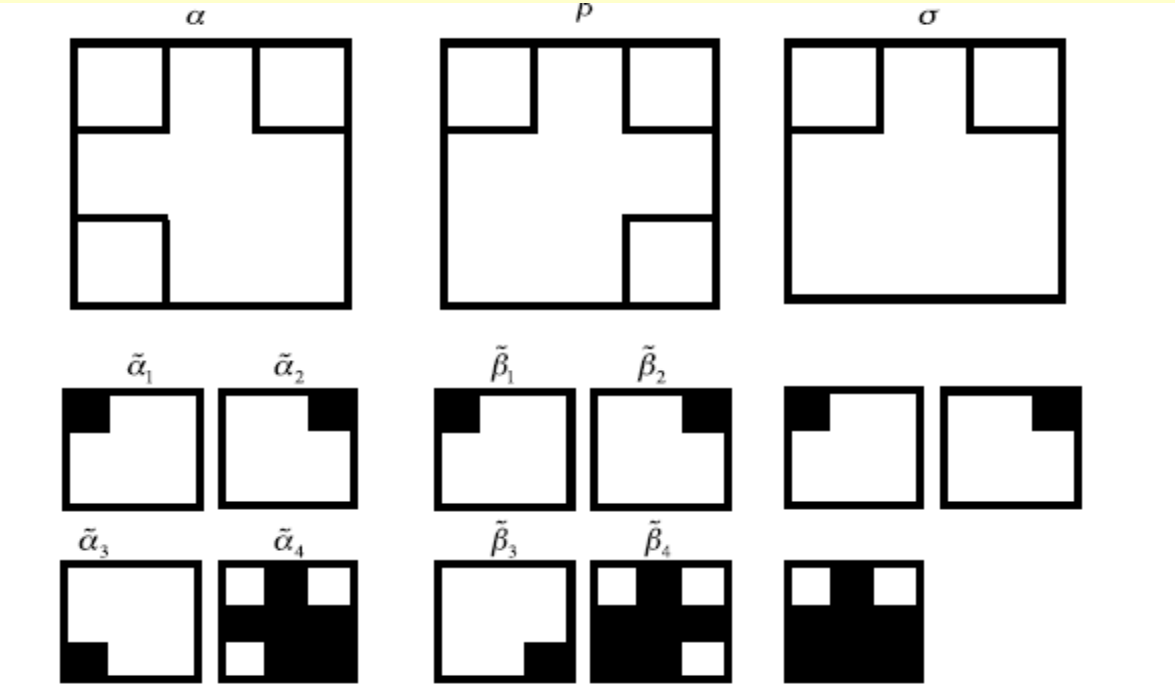
(Same procedure for **β : l**et $\boldsymbol{\beta'_j}$ be the surviving factors of **β**)

iii – Define the reduced partitions **α'** and **β'** as

$$\boldsymbol{\alpha' = V_k\, \alpha'_k} \quad \text{and} \quad \boldsymbol{\beta' = V_1^N\, \beta'_j}$$

It follows: $\boldsymbol{d_R(\alpha',\beta') \geq d_R(\alpha,\beta)}$, **i.e. amplification**

Elementary Factors and Intersection

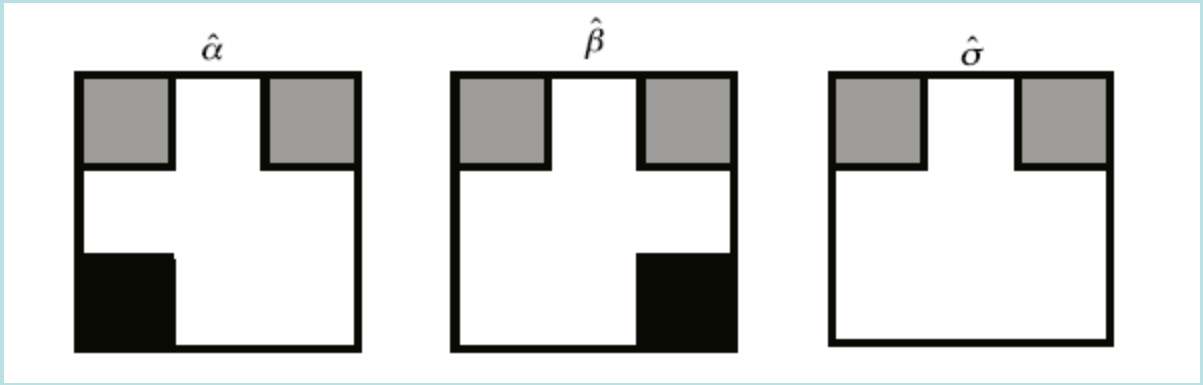


(From E.Agliari, M.Casartelli, E.Vivo, J.Stat.Mech.(2010)).

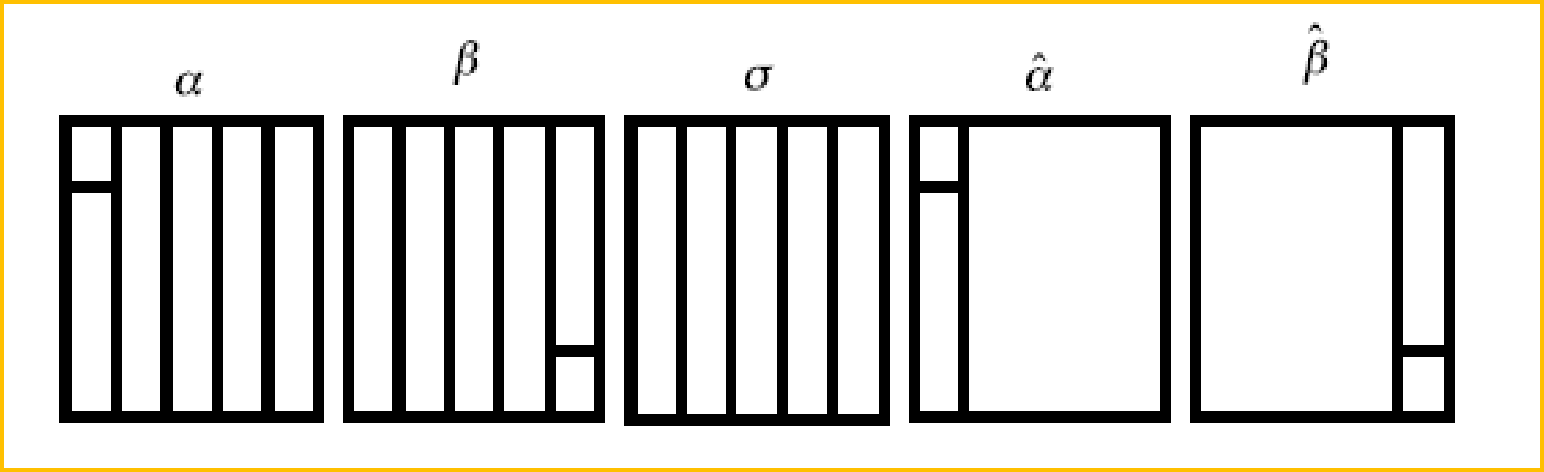
And then….

The result of the reduction process;

(note that there appear non-connected atoms)



Another example



The method can be applied to the evolution in arbitrary Configuration Spaces, e.g. for **Ising Models** on graphs, **Sandpiles**, or **Cellular Automata** in general.
Configurations must be projected in the Partitions Space of the supporting structure (the set of site labels)

Ingredients:

X = (**M**, $\mathscr{M}$, **μ**)

An alphabet $\mathbb{K}$

The Configuration Space ***C*(M)**

The Partition Space $\mathscr{Z}$**(M)** obtained from: ***C*(M)** ⟶ $\mathscr{Z}$**(M)**

There are partial algebraic results, valid also on arbitrary graphs

Example :

The necessary and sufficient condition in order to have an amplification of the Rohlin's distance is that, in the intersection $\sigma = \alpha \wedge \beta$, there exist at least two simple atoms from the same partition and at least one atom which is composed for a partition and simple for the other one.

(From E.Agliari, M.Casartelli, E.Vivo, J.Stat.Mech. 2010)

Partitions originated from connected homogeneous subsets can play a special role. In such cases, the geometric structure of the support can allow for an alternative approach to the reduction process, with a different choice of the elementary factors.

Finite one-dimensional strings of characters, such as

CGTUGTTGUUUCT

suggest the right choice of the elementary factors, exploiting the natural order of the site labels {*1,2,3,…,L*}

The projection Φ : $\mathcal{C} \rightarrow \mathcal{Z}$

From

a = $(a_1, a_2, a_3, \ldots, a_L)$ = (*F F G G G H F F F F*)

to

α = **Φ**(**a**) = (A_1, A_2, A_3, A_4) = ((1 2),(3 4 5),(6),(7 8 9 10))

A probability measure μ on the subset-algebra $\mathcal{M}$ of **M** is given by the normalized number of sites in each subset:

in the example: probabilities associated to **α** are (2/10, 3/10, 1/10, 4/10).

Important: Such a probability is exclusively based on the structure of the text. Remember that homogeneous segments are meaningless in themselves.

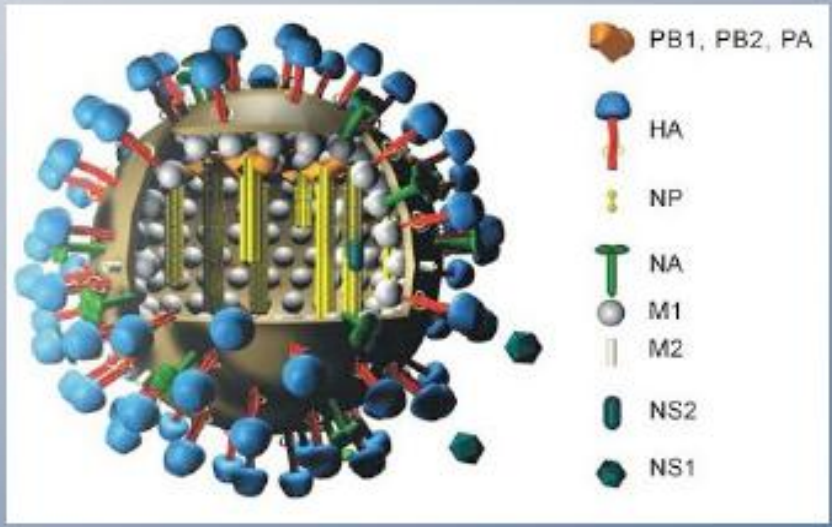
The one-dimensional strings peculiarities naturally lead to the analysis of DNA and RNA sequences as “texts”, provided that:
-They are homogeneous (same length)
- Time evolution is a meaningful problem for them.

There exist several biological contexts where such features apply and our method can work.

We have chosen one….

Influenza A H3N2

Influenza A Virus: a rapidly evolving disease



120 nm, 8 proteins

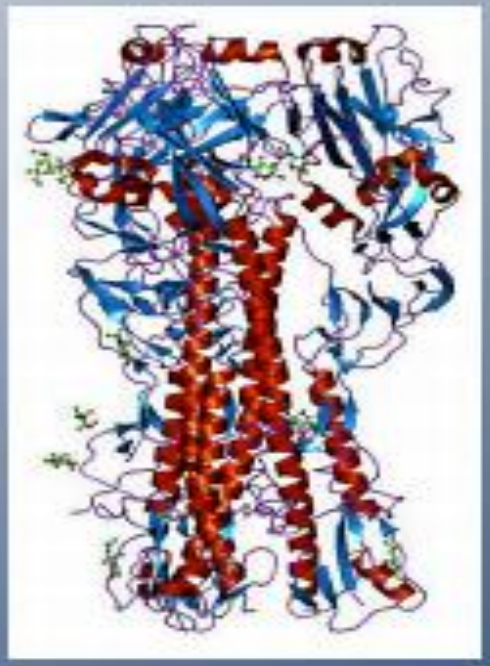
1. Polymerase B2 protein (PB2)
2. Polymerase B1 protein (PB1)
3. Polymerase A protein (PA)
4. Hemagglutinin (HA or H)
5. Nucleocapsid protein (NP)
6. Neuraminidase (NA or N)
7. Matrix protein (M): M1 constructs the matrix; and in influenza A viruses only, M2 acts as an ion channel pump to lower or maintain the pH of the endosome
8. Non-structural protein (NS); the function of NS2 is hypothetical

the whole genome evolve but...

Focus on Hemagglutinin: why?

Influenza A Virus: a rapidly evolving disease

The most important protein involved in recognition by the immune systems is the Hemagglutinin (from the protein's ability to cause red blood cells to clump together, i. e. agglutinate)



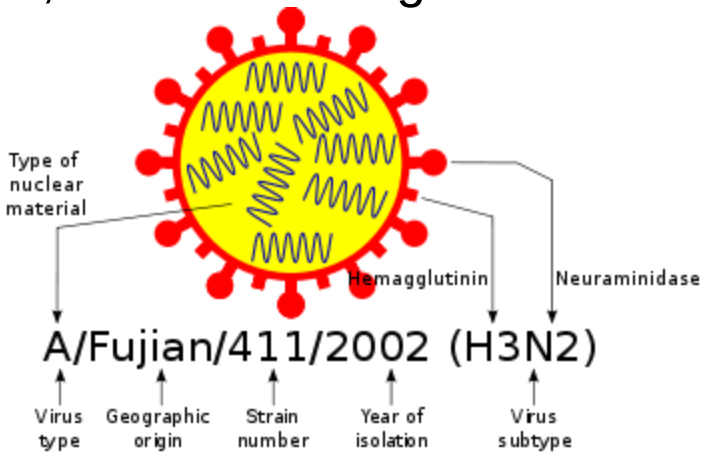
In particular, changes in the Hemagglutinin can be very effective and they can allow the virus to escape from the immune system, while preserving its fitness.

The HA is a proteins, so changes occurs in the sequences of its amino acid, during replication (error rate: 1 every 10^4 bases).

an example:

From Public Databases….

A virus is isolated and its Hemagglutinin is sequenced and put in public databases with a label indicating type, place and time
The H squence is a word of ~ 400 letters taken from an alphabet of 20 letters, each indicating one amminoacid.



Influenza A Virus: a rapidly evolving disease

Positions from 61 till 120

```
          Consensus Sequence          555T6GICD5PHQILDGENCTLIDALLGDPQCDGFQNKKWDLFVERSKAY5NCYPYDVPD
ABQ10126  A/Sydney/5/1997(H3N2)      .....R......R..................K........E....................
CAC37007  A/Sydney/5/1997(H3N2)      .....R......R..................K........E....................
AAT08602  A/Moscow/10/1999(H3N2)     .....R.........................K........E....................
ABE73115  A/Moscow/10/1999(H3N2)     .....R.........................K........E....................
ABF21273  A/Panama/2007/1999(H3N2)   .....R.........................K........E....................
ABE73114  A/Panama/2007/1999(H3N2)   .....R.........................K........E....................
ACF54554  A/Panama/2007/1999(H3N2)   .....R.........................K........E....................
```

Positions from 181 till 240

```
          Consensus sequence          NVTMPNNEKFDKLYIWGVHHPGTDNDQISLVAQASGRITVSTKRSQQTVIPNIGSRPRVR
ABQ10136  A/Sydney/5/1997(H3N2)      ........D..............S..S..T.I.......V.....................W..
CAC37007  A/Sydney/5/1997(H3N2)      ........D..............S..S..T.I.......V.....................W..
AAT08602  A/Moscow/10/1999(H3N2)     ........D..............S..S..T...T.....V.....................W..
ABE73115  A/Moscow/10/1999(H3N2)     ........D..............S..SV.T.V.V.....V.....................W..
ABF21273  A/Panama/2007/1999(H3N2)   ........................S..S....I.......V.....................S.W..
ABE73114  A/Panama/2007/1999(H3N2)   ........................S..S....I.......V.....................I.W..
ACF54554  A/Panama/2007/1999(H3N2)   ........................S..S....I.......V.....................S.W..
```

n.b. : they are of the same length

Evolution and metric properties

- How different are the sequences?
  Which mutations are relevant, causing reinfection?

- Can one characterize the evolution, generated by interaction with the immune system? It is continuous or punctuated? It is a “drift”?Is it possible to identify regularities and use them to predict the next prevailing strain?

-The general idea is that the information encoded in the sequence is strictly related to the properties of the corresponding biological structure

Assume that

a, b, c, …. are sequences of fixed length L :

$$\mathbf{a} = (a_1, a_2, a_3, \ldots, a_L) ,$$

(where a_K is in $\mathbb{K}$**,** the alphabet of aminoacids)

Their set is the

Configuration (or *State*) *Space* $\mathscr{C} \equiv \mathscr{C}(\mathbf{M})$

over **M**, the set of site labels (1,2,3, …., L)

Previous and current approaches…

An obvious (and widely used) metrics in $\mathcal{C}$ is d_H **(a, b),**

the Hamming Distance :

$$\mathbf{d_H(a, b) = \sum_k (1 - \delta(a_k, b_k))}$$

(It simply counts the number of sites with different symbols, ignoring correlations…)

Recent improvements:

- Inequivalent labels (only epitopes are considered…)
  (Deem, Vaccine 2010, PRL 2007)
- Weights on couples of symbols, from frequencies in historical series (Miyata metrics)

Another (biochemical) approach…

Calculating distances between viral strains

On the opposite side, there are "antigenic" metrics. The most famous and widely used (i.e. by the WHO), is based on HI (hemagglutination inhibition) assays. It is an "antigenic" distance, calculated on the ability of the animal antisera raised against one virus to stop the agglutination of red blood cells caused by another virus.

- It is certainly true, but it is very imprecise (distance based on log of the HI assays concentration, and depend on many external factors) (Smith, Lapedes et al, Science 2004). According to this metrics, the evolution is very discontinuos!

"Gradual genetic but punctuated antigenic change"

Rohlin's Distance for the Influenza Amminoacids Sequences

We need a projection $\Phi : \mathcal{C} \rightarrow \mathcal{Z}$

The partition atoms are individuated by homogeneous (i.e. same character) segments in the sequence:

i.e. from

$\mathbf{a} = (a_1, a_2, a_3, \ldots, a_L) = (F\ F\ G\ G\ G\ H\ F\ F\ F\ F)$

to

$\boldsymbol{\alpha} = \boldsymbol{\Phi}(\mathbf{a}) = \{A_1, A_2, A_3, A_4\} = \{(1\ 2),(3\ 4\ 5),(6),(7\ 8\ 9\ 10)\}$

In this case, the convenient elementary factors of $\mathcal{E}(\boldsymbol{\alpha})$ are :

$$\boldsymbol{\alpha}_k = \{(A_1 U A_2 \ldots U A_k)\ ,\ (A_{k+1} U \ldots U A_N)\}$$

Probability

A probability measure **μ** on the subset-algebra $\mathcal{M}$ of **M** is given by the normalized number of sites in each subset: in the previous example: probabilities associated to **α** are

(2/10, 3/10, 1/10, 4/10).

Important: Such a probability is exclusively based on the structure of the text. Remember that homogeneous segments are biologically meaningless in themselves.

This is a **"degre zero"** approach (true **black box** analysis). (Alternative possibilities are in progress…)

Examples (with $L = 10$)

$\mathbf{a}=(a_1,a_2, a_3, \ldots,a_L) = (F\ F\ G\ G\ G\ H\ F\ F\ F\ F)$

$\mathbf{b}=(b_1,b_2, b_3, \ldots,b_L) = (H\ H\ H\ F\ F\ G\ G\ G\ G\ G)$

$\boldsymbol{\alpha} = \Phi(\mathbf{a}) = ((1\ 2),(3\ 4\ 5),(6),(7\ 8\ 9\ 10)) \rightarrow (1,3,6,7)$

$\boldsymbol{\beta} = \Phi(\mathbf{b}) = ((1\ 2\ 3),(4\ 5),(6\ 7\ 8\ 9\ 10)) \rightarrow (1,4,6)$

“**reduction process π**” on couples of partitions for alphabetical strings

Remembering the process **π**: **(α,β)** → **(α’,β’)**
which amplifies the Rohlin’s Distance by erasing as far as possible the common sub-partitions, only in the case of character strings, with the left border representation above, the reduction process of partitions corresponds to the

cancellation of common *left* borders $k > 1$.

$\alpha = \Phi(\mathbf{a}) \rightarrow (1,3,6,7)$; prob=(2/10, 3/10, 1/10, 4/10)

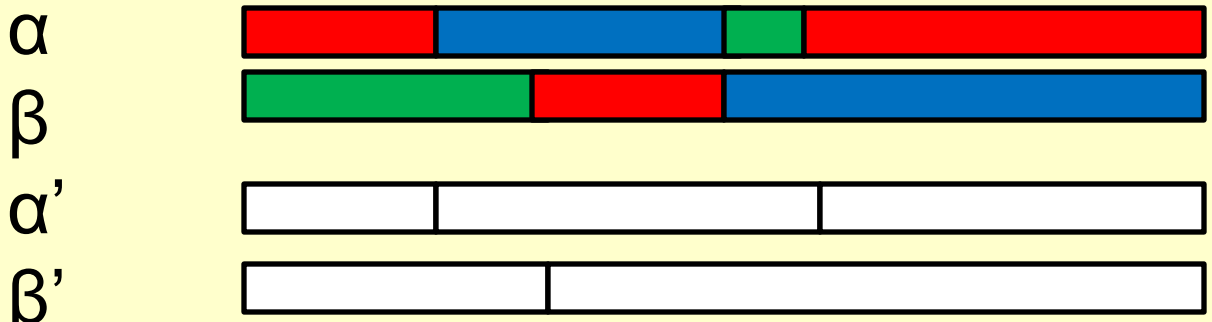
$\beta = \Phi(\mathbf{b}) \rightarrow (1,4,6)$; prob=(3/10, 2/10, 5/10)

$\alpha' = (1,3,7)$; prob = (2/10, 4/10, 4/10)

$\beta' = (1,4)$; prob = (3/10, 7/10)

$d_R(\alpha,\beta)=0.1541097$; $d_R(\alpha',\beta')=0.38822625$

warning: reduced partitions do not correspond to real sequences!



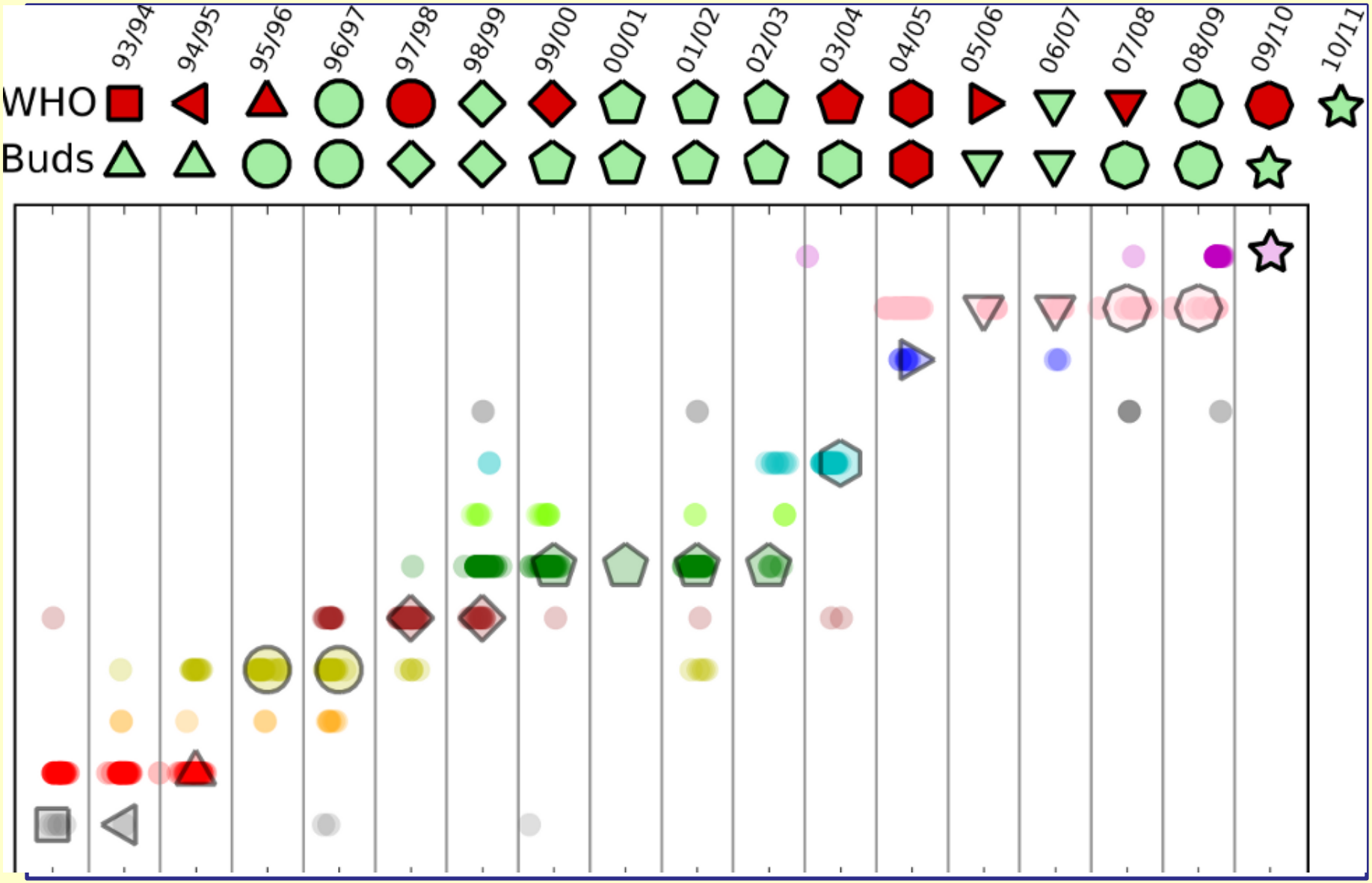
Matrices of Distances and Clustering

$\mathbf{H}_{ik} = d_H(\mathbf{a}_i, \mathbf{a}_k)$
$\mathbf{R}_{ik} = d_R(\alpha_i, \alpha_k)$
$\mathbf{R'}_{ik} = d_R(\alpha'_i, \alpha'_k)$ (this is the relevant one!)

These matrices regard the whole set of N=824 sequences, each of them with its sampling date (including the WHO's ones)

A clustering method is applied (**complete linkage hierarchical algorithm**) with the number *p* of clusters as an external parameter.

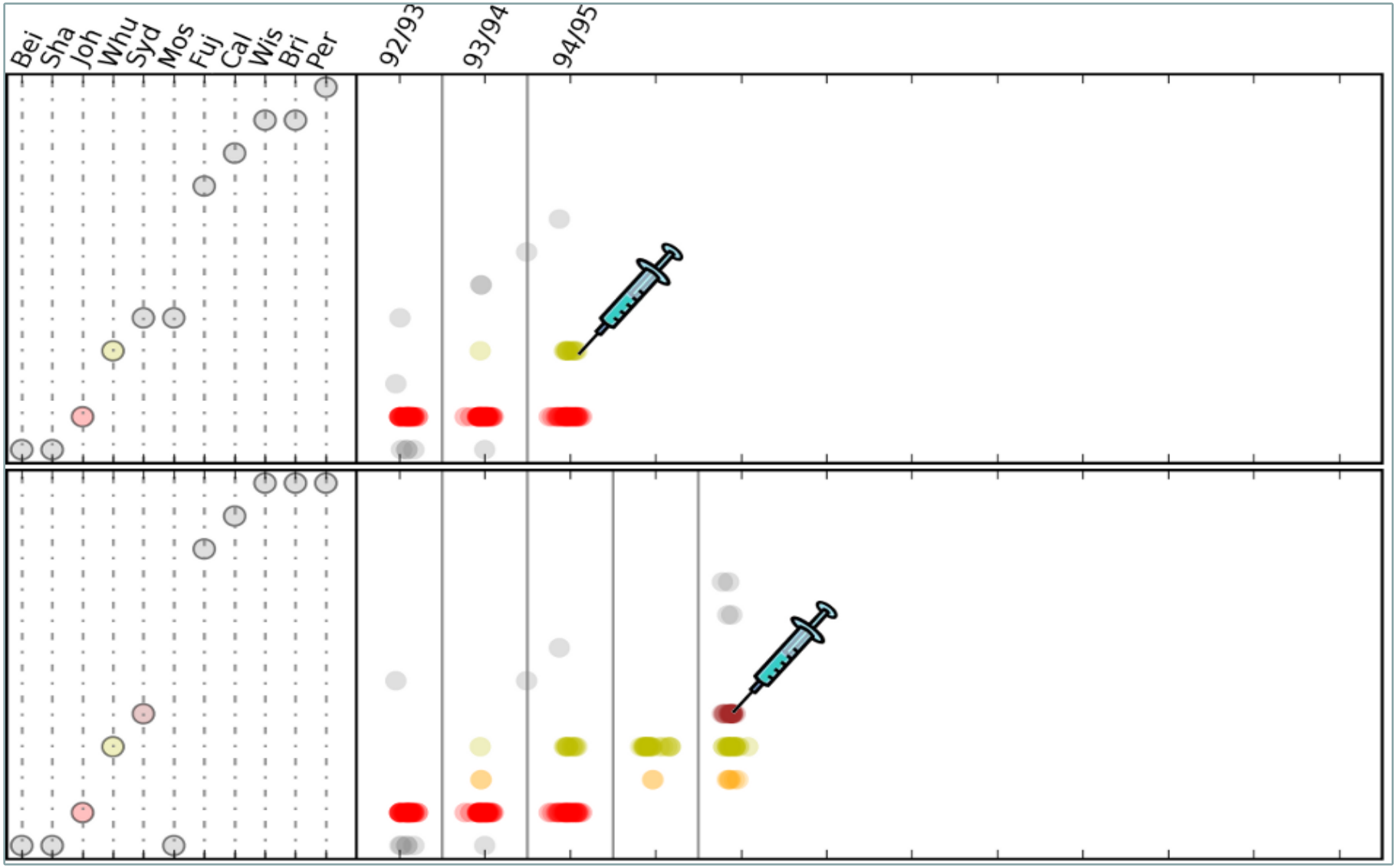
Clustering, as they appear vs. time



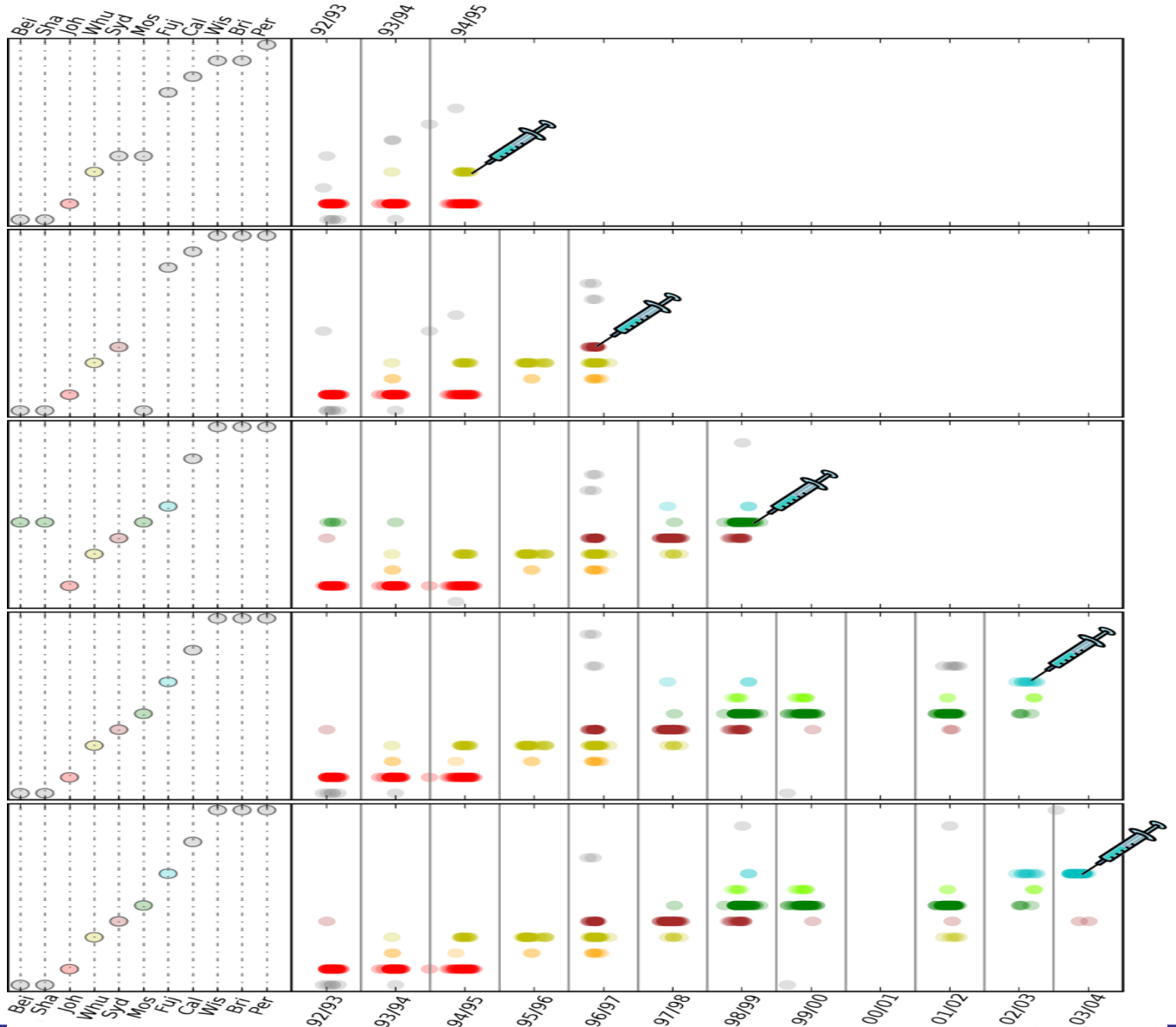
Looking for the next prevailing strain: the Buds

From the clustering analysis it is possible to extract
a criterion to mark the "winning clusters"
which will become dominant the next year: it appears that
Emergent buds, when clearly marked,
are the winning one in the next season.

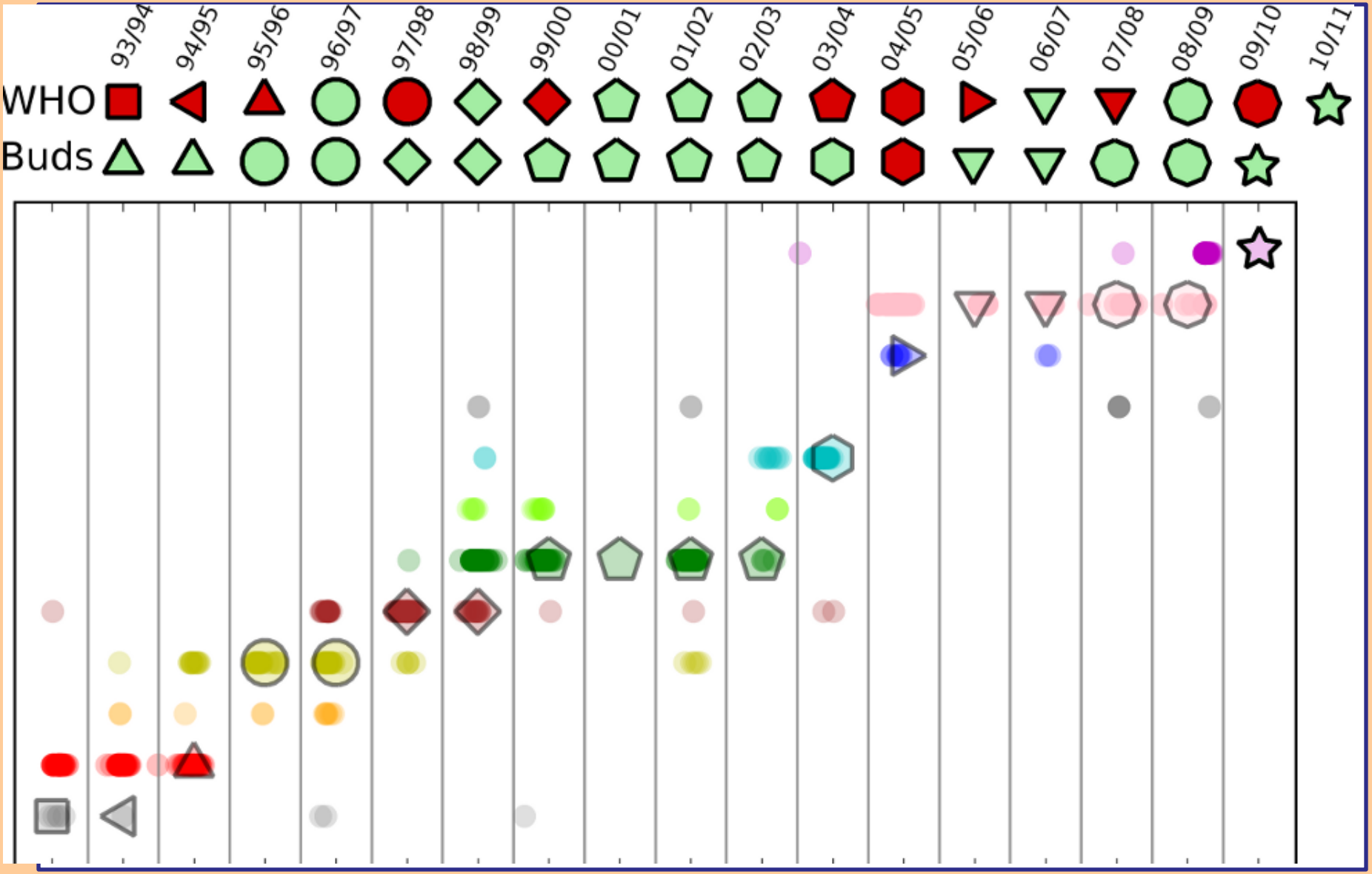
Variable Time Window in the Data Sampling



Again and again…..



Looking again…



Robustness of the results

With respect to the considered time window (ok)

With respect to the clustering parameter *p* *(ok)*

With respect to label permutations (*ok)*

With respect to the type of influenza and the geographical area (in progress, but *ok…)*

Example: how to choose the *p* parameter in clustering

- If $n_1, n_2, \ldots n_p$, are the number of sequences contained in the p clusters out of N, define the weights

$$w_k = \frac{n_k}{N} \qquad (k = 1, 2, \ldots p)$$

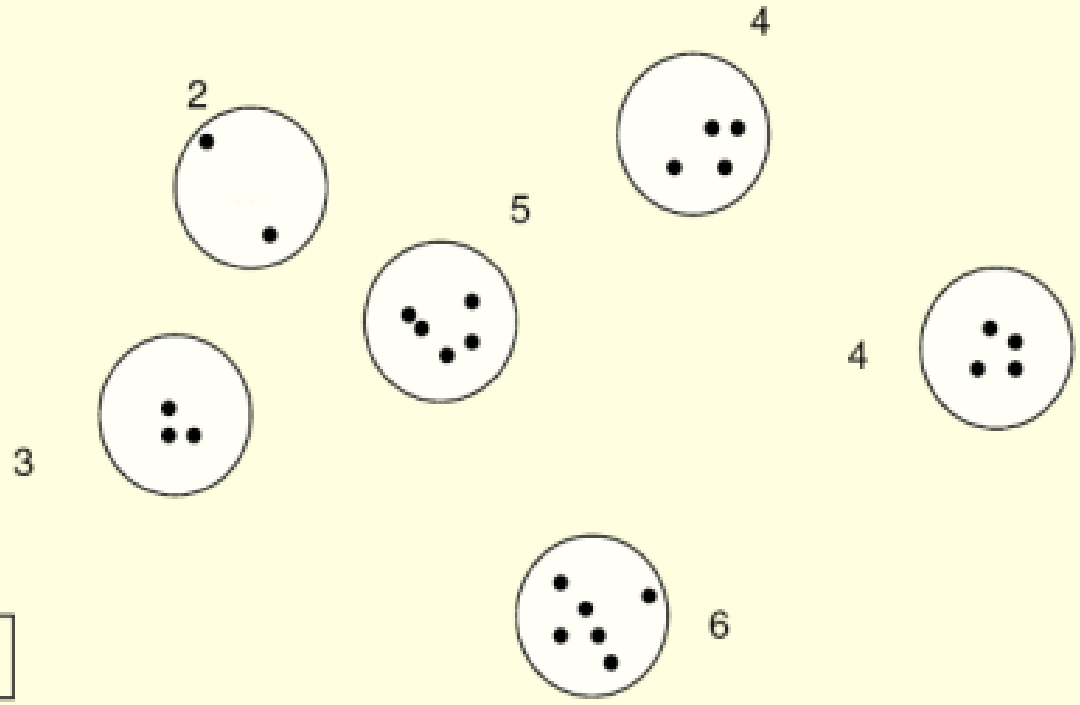
- consider the Shannon entropy of this distribution

$$\underline{w} = (w_1, w_2, \ldots w_p)$$

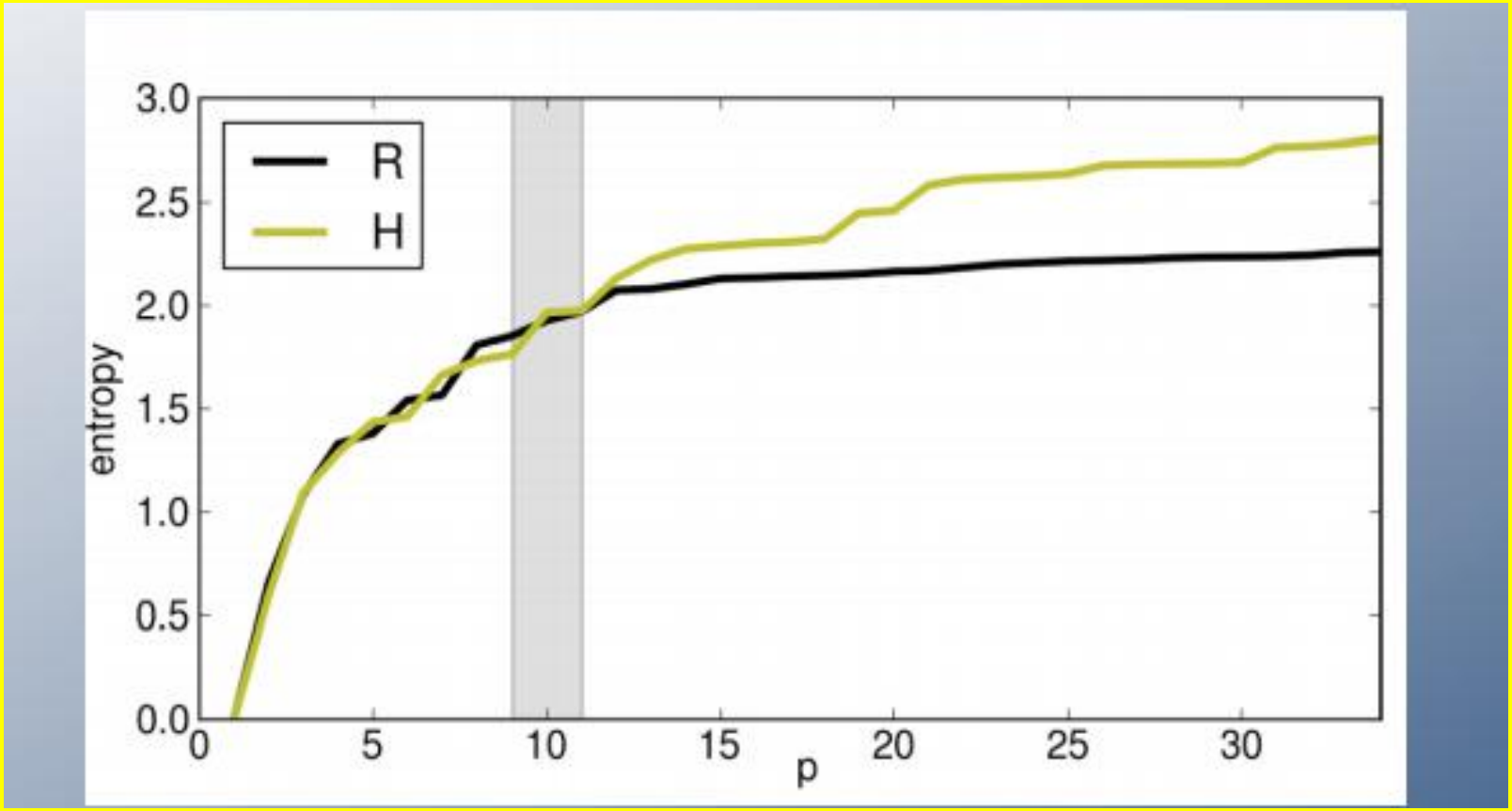
$$H(\underline{w}) = -\sum_k w_k \ln w_k$$

as a function of p. As p grows, due to possible splitting, H is not decreasing. If it is stable it means that clusters do not split.

W = (2/24, 3/24, 5/24, 4/24, 4/24, 6/24)



$p = 6$



Our experimental p coincides with the number of strains indicated by the WHO in the same period

Influenza A evolution as seen by Rohlin metrics:

Other datasets features the same scenario
and the same effectiveness of the bud criterion:

- H1N1 - North America: (apart from the 2009 pandemy: no one can predict antigenic shifts!). Interesting features emerges, instabilities of the post-pandemic period

- New Zealand H3N2

This is the first step in this direction:

R.B., R.Scalco, M. Casartelli, to appear on PLoS One (2011)