

**BMM 1**

**Presentation**

Statistiques Multivariées pour la Bioinformatique

**Susan Holmes, Stanford, [susan@stat.stanford.edu](mailto:susan@stat.stanford.edu)**

## Spécificités de cette classe

- Les données génétiques sont discrètes: comptage .
- En général les données ne sont pas indépendantes.
- Grands ensembles de données, méthodes comme le data mining.
- On a besoin d'utiliser des interfaces entre bases de données et logiciels statistiques.
- Les paramètres sont non-standards.

## Buts de ce module

Apprendre les outils statistiques spécifiques aux données génétiques, proteomiques modernes.

- Variables aléatoires discrètes.  
(Binomiale, Multinomiale, Poisson, Dirichlet,)
- Simulation de Monte Carlo.
- Methodes Multivariees pour donnees quantitatives, exploratoires.

- Methodes Multivariees pour donnees a expliquer.
- Lissage

## Apprendre des outils statistiques spécifiquement ciblés pour les grands données

- Valeurs extrêmes. (maximum , minimum)
- Analyses Multivariées (ACP, DVS, AC, AD, Classification Hierarchique).
- EM, méthodes Bayésiennes, stabilisation de la variance, MV.
- Regression non paramétrique (lissage).

## Connaître des outils logiciels d'analyses de données genomiques/proteomiques

- 
- Methodes de Projection
- Methodes Exploratoires for tableaux de contingence.
- Methodes de Discrimination.
- Méthodes de Classification.
- Visualization de données multivariées.

- Analyse de puces d'ADN.
- Simulation de Données pour vérifications de modèles.

## Un monde de variabilité

La biologie est encore plus difficile à comprimer en des principes simples que la physique. Tout est variabilité.

C'est la variation qui a permis l'évolution et c'est cette variabilité qui assure la robustesse de complexes systèmes biologiques. Cette variabilité agit en règle plutôt qu'en exception en biologie.

Statistiques et probabilité agissent comme des outils essentiels dans la décomposition de signaux présents dans les données génétiques.

# Particularités des données genomiques

- Sequences genetiques: les données sont discrettes, ou bien binaires ou avec un nombre reduit de categories (A, C, G, T), sous forme de frequences ou de tableaux de contingence.

Exemples:

Des transition entre etats

	followed by		
first	AT	AG	AT
CG	34	14	33

CT	12	10	10
CA	11	17	09

## Donnee de Phenotype

eyes	Black	Brunette	Red	Blonde
Brown	68	20	15	5
Blue	119	84	54	29
Hazel	26	17	14	14
Green	7	94	10	16

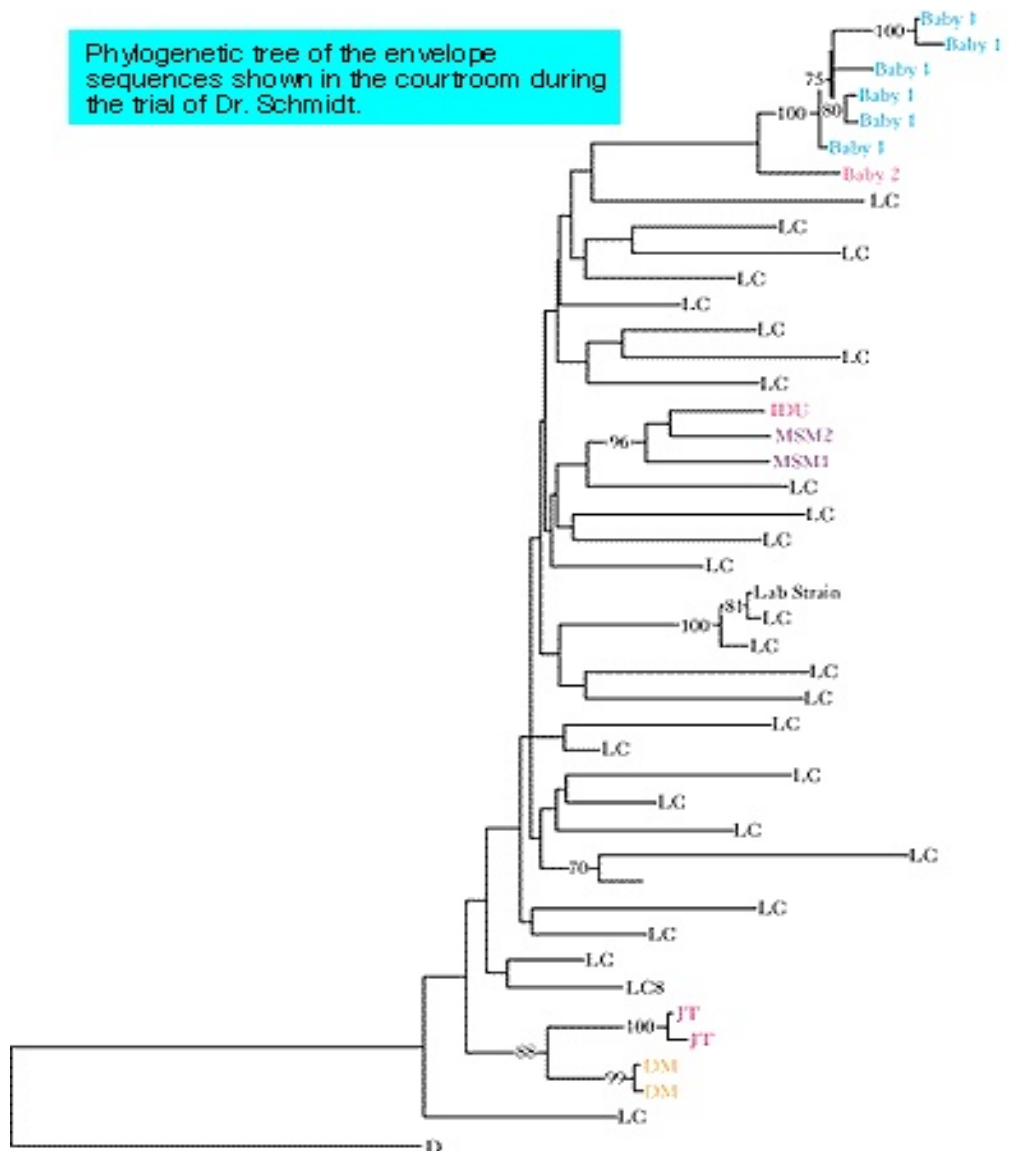
- Independence des observations ou variables n'est pas la norme, en general les données se caracterisent par un bon niveau de dependence (donc les chaines de Markov se montrent une outil de choix).
- De tres grands ensembles de données qui sont beaucoup plus

communsque dans dans n'importe quel autre domaine.

- We will need to interface statistical procedures with the large genetic database searches (the glue can be languages such as perl, R,python).

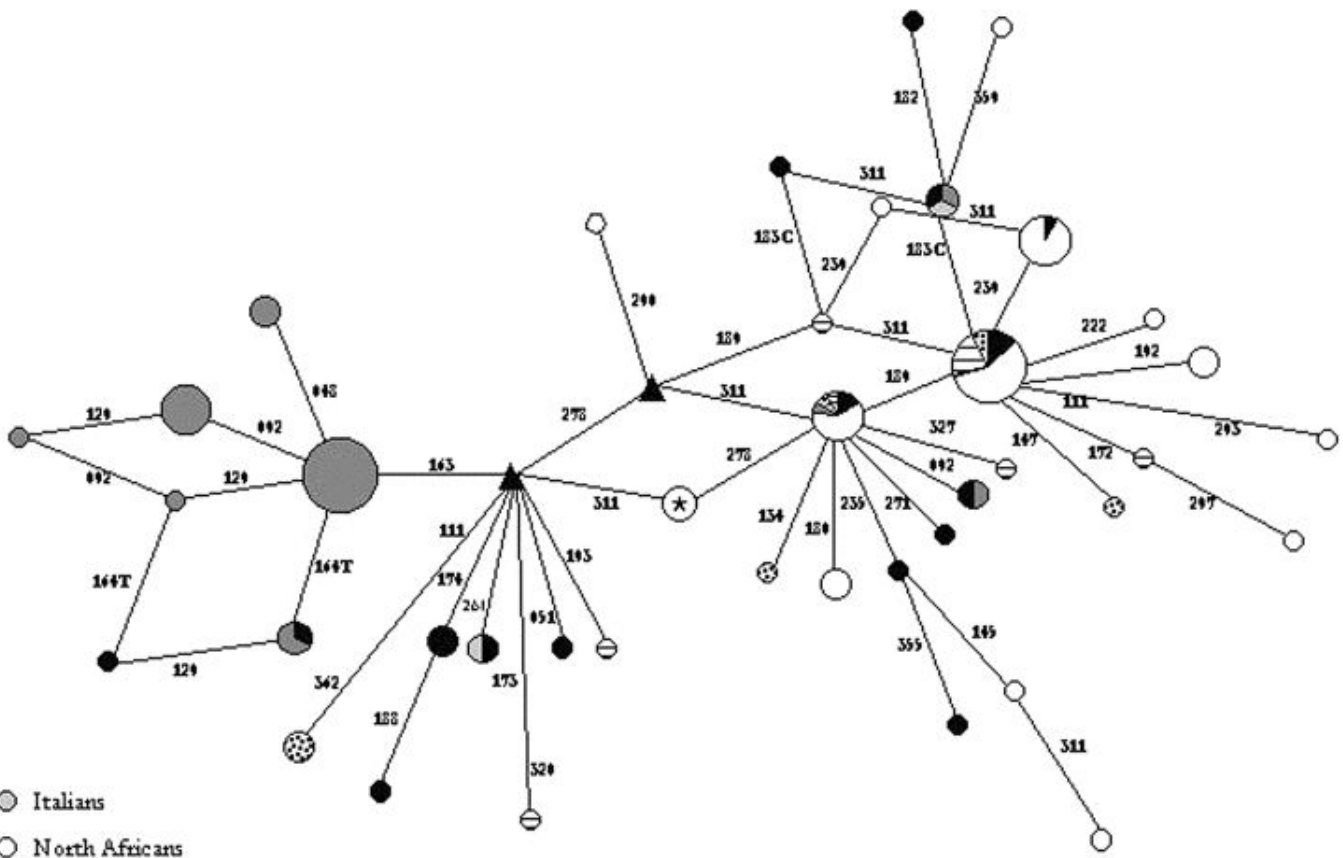
- Les paramètres auxquels on s'intéresse sont non standards, pas seulement des vecteurs, mais des arbres, des

Phylogenetic tree of the envelope sequences shown in the courtroom during the trial of Dr. Schmidt.



permutations, des reseaux

10%

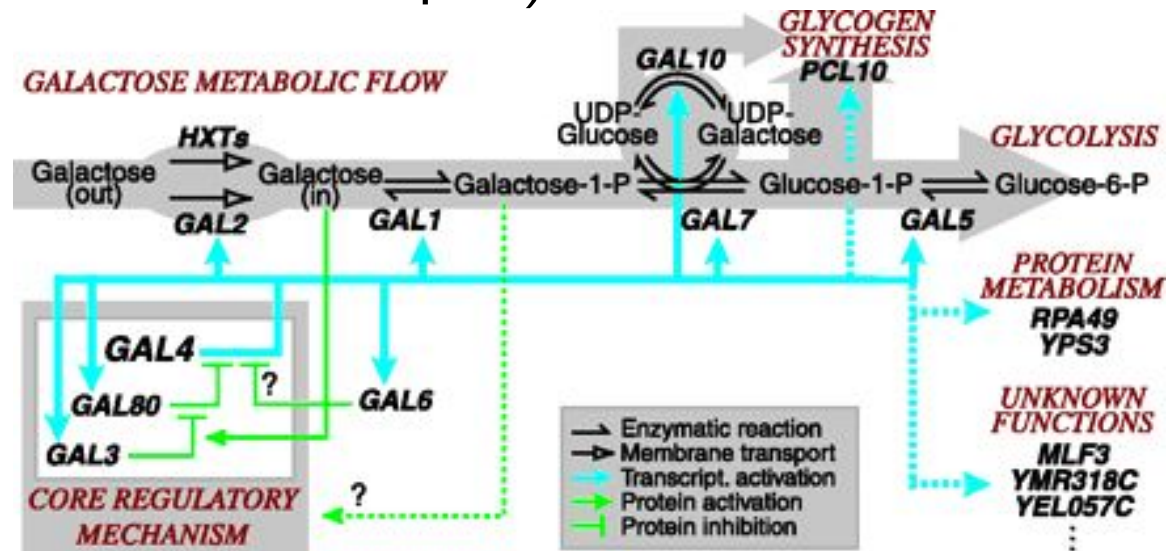


- Italians
- North Africans
- ⊖ South Africans
- Iberians
- Canarians
- ⊗ Middle East

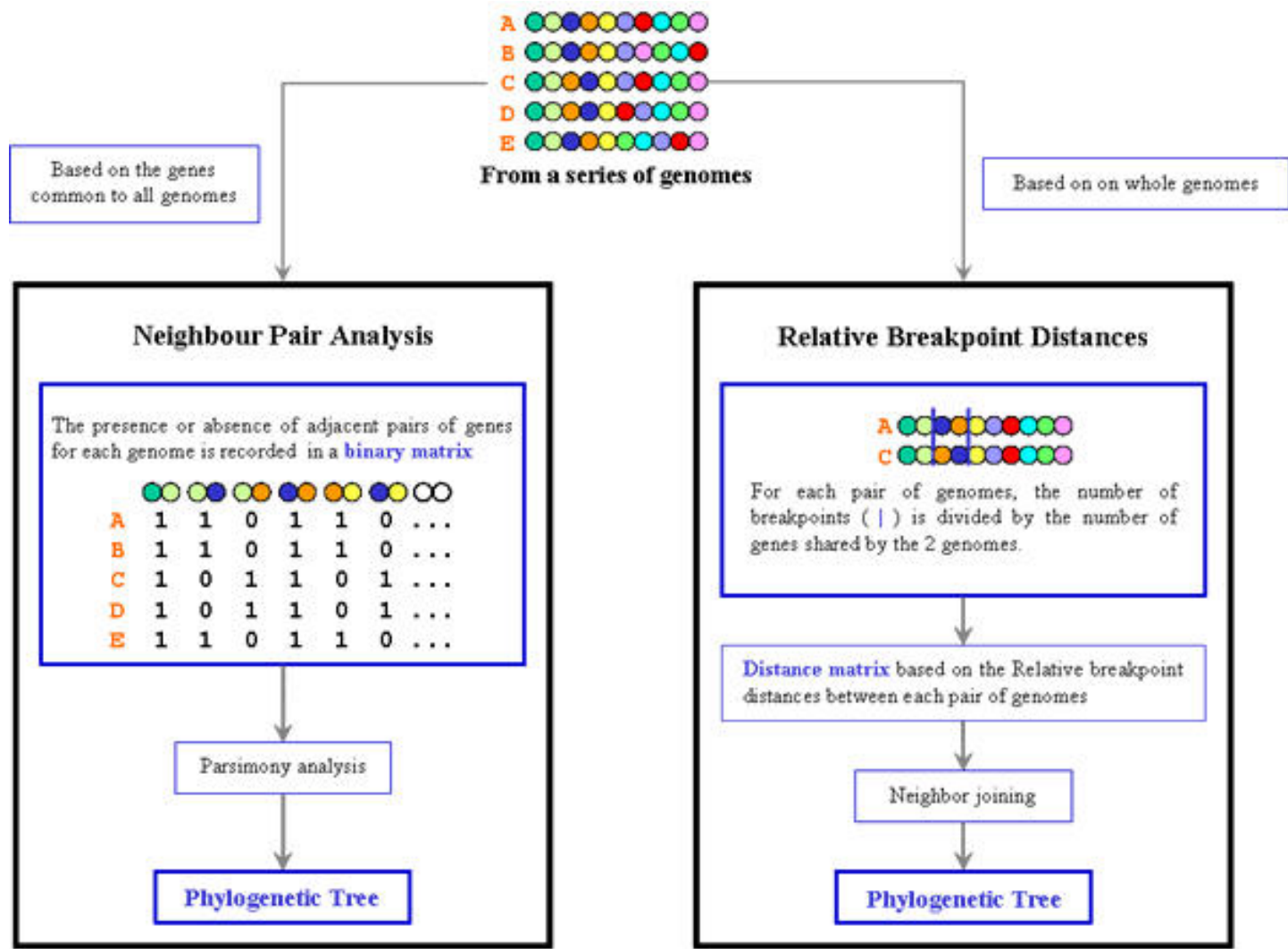
(les genes travaillent de consort et c'est important de comprendre comment ils interagissent dans des reseaux de transcription)



ou des reseaux metaboliques)



(permutations).



## Exemple

A la recherche d'epitopes

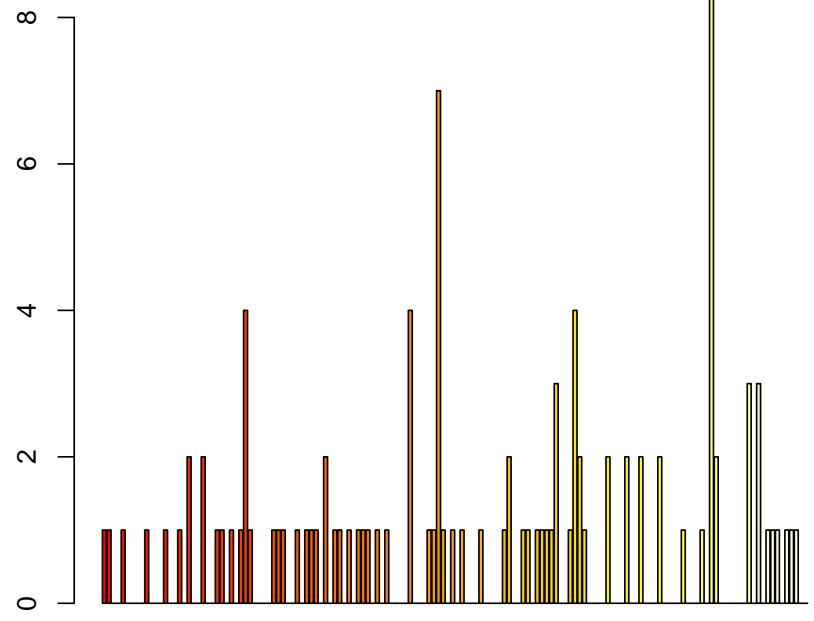
Epitope: *Une portion specifique d'antigen macromoleculaire a laquelle se lie un anticorps. Dans le cas d'un proteine antigen reconnu par un lymphocyte (T cell) l'epitope ou le determinant est une portion de peptide qui se lie avec la molecule MHC Major Histocompatibility Complex (MHC) en reconnaissance du T cell receptor (TCR)*

Si on dispose d'un test chimique pour la reaction allergique qui produit une reponse 0/1, mais il y a un bruit de fond avec

des 1 apparaissant meme quand ils ne devraient pas avec une petite probabilité  $p_0 = \frac{1}{100}$ ,

Si la sequence de proteines testée provoque une réaction allergique, (nous appellerons la sequence une epitope), la probabilité de voir une reaction sera proche du nombre de personnes qui presentent cette allergie particuliere. par exemple  $p = \frac{10}{100}$ .

On teste 150 protein sequences (150 fenetres mobiles sur 50 personnes), on obtient comme nombre de 1:



Modele probabilistique simple: Fond Poisson iid au taux  $50/100 = .5 = \lambda$ . Le maximum ici est 9: avec quelle probabilite va-t-on rencontrer une valeur aussi grande que 9 dans un echantillon de Poisson(0.5)?

Here we will use **extreme value theory**: the distribution of the order statistic  $x_{(n)} = \max\{x_1, x_2, x_3, \dots, x_{150}\}$ .

Quelles sont les chances de voir un maximum de 9?

$$P(x_{(n)} \geq 9) = 1 - P(x_{(n)} < 9) = 1 - \prod_{i=1}^n P(x_i < 9)$$

en supposant que les 150 soient independent

$$\prod_{i=1}^n P(x_i < 9) = \left( \sum_{k=0}^8 \frac{e^{-\lambda} \lambda^k}{k!} \right)^n = \left( 1 - \sum_{k=9}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} \right)^n$$

identiquement Poisson

Appelons  $\epsilon = \sum_{k=9}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} = \mathbf{P}(x_i \geq 9)$ .

De quelle taille est  $\epsilon$ ?

R

```
> sum(dpois(0:5,0.5))
```

```
[1] 0.9999858
```

```
> ppois(5,0.5)
```

```
[1] 0.9999858
```

```
> sum(dpois(0:8,0.5))
```

```
[1] 1
```

$$(1 - \epsilon)^n = \exp(n \log(1 - \epsilon)) = \exp(-n\epsilon) = e^{-nP(x_i \geq 9)}$$

Comment conclure que cette bosse correspond a une epitope?

```
> epsilon=1-ppois(8,0.5)
```

```
> epsilon
```

```
[1] 3.43549e-09
```

```
> 150*epsilon
```

```
[1] 5.153236e-07
```

```
> exp(-150*epsilon)
```

```
[1] 0.9999995
```

```
> 1- exp(-150*epsilon)
```

```
[1] 5.153234e-07
```

$$P(x_{(n)} \geq 9) = 1 - e^{-nP(x_i \geq 9)} = 1 - 0.9999995 = 5.10^{-7}$$

This is a very small probability.

Outils des probabilistes les plus utiles en bioinformatiques:  
les lois de probabilités pour les variables discrètes: binomiales,  
Beta, multinomiales, Dirichlet, Poisson,  $\chi^2$ .

Simulations de Monte Carlo sont utiles dans les situations complexes.

References (en anglais): Eddy, Krogh, Mitchison  
Probabilistic Analysis of sequence data.  
Ross Pitman and Grinstead and Snell

# Qu'est-ce que la probabilité?

Ce sont les calculs analytiques, mathématiques qui nous permettent d'aller un modèle hypothétique, comme les modèles d'évolution vers la probabilité d'un événement, un ensemble de données par exemple:

Attention: au grain de ble...

En mathématiques, la théorie des probabilités permet de modéliser l'incertitude et le bruit et d'évaluer leur effet sur les données.

# Qu'est-ce que la Statistique?

Sens inverse: on va aller du reel , des donnees au modele qui est le plus vraisemblable, le plus plausible mais aussi le plus simple.

Principe de Parcimonie (Ockham's razor).

Les statistiques modernes sont basés sur l'analyse des données pour ce qu'elles sont et souvent commence sans modele: on dit des analyses modernes que ce sont des moteurs de generation d'hypotheses, au contraire de ce qui se passait avant.

Les données sont massives (high throughput) et de très hautes dimension.

## Hardest part: curse of dimensionality.

La complexité et la grande dimension des phénomènes biologiques se heurtent à la rarification qui survient en haute dimension: on appelle cela le fléau de la dimensionalité.

On ne peut pas étudier les variables une à une car on perd toute l'information due à la dépendance, par exemple on verra que les données d'expression des gènes sont très corrélées on perd l'information sur les réseaux si on se limite à des études à une ou deux dimensions.

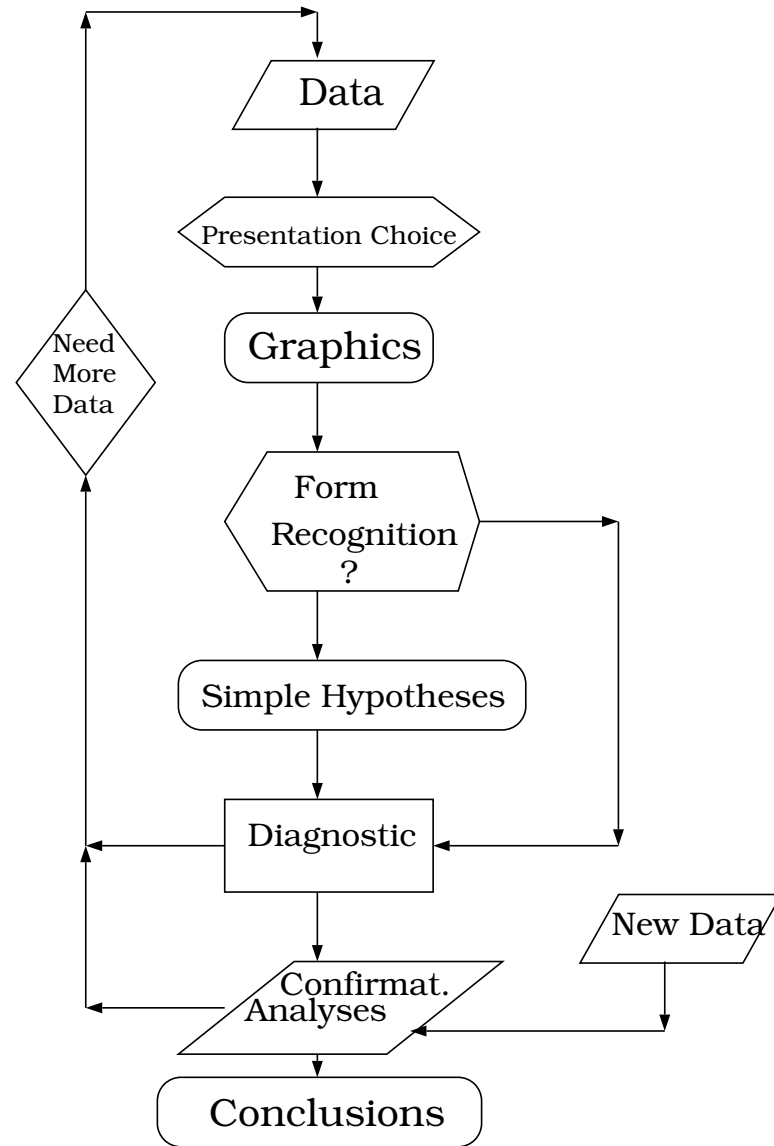
On va maintenant faire un survol des méthodes multivariées

utilisees en bioinformatique :ACP, CA, MDS , Classification hierarchique ou automatique.

Sans modele il vaut de multiples facons de visulaizer et simplifier les donnees, on cherch en fait des variables sous-jacentes cachees: ce sont les motivations des methodes multivariees.

# Statistiques ne sont plus que p-values

La caricature des test d'hypotheses sur des variables unidimensionnelles datent des annees 50, maintenant la statistique utilise l'informatique de facon intensive et remplace les calculs analytiques par les simulations.



# Web References

- Probability Distributions
- Probability by Surprise
- Statistics of Sequence Similarity Scores
- Genetics Glossary for the layperson