# Evolution and Structure of Biomolecules

Julien Dutheil[1]

<jdutheil@daimi.au.dk>

[1]BiRC – Bioinformatics Research Center,
University of Århus

http://birc.au.dk/~jdutheil/Teaching/

February 2008

# Introduction

## Which molecules are we interested in?

- *Evolving molecules*: DNA, RNA, Proteins
- Several structures are resolved: we have the set of coordinates for almost all atom positions
- There are several levels of structure organization

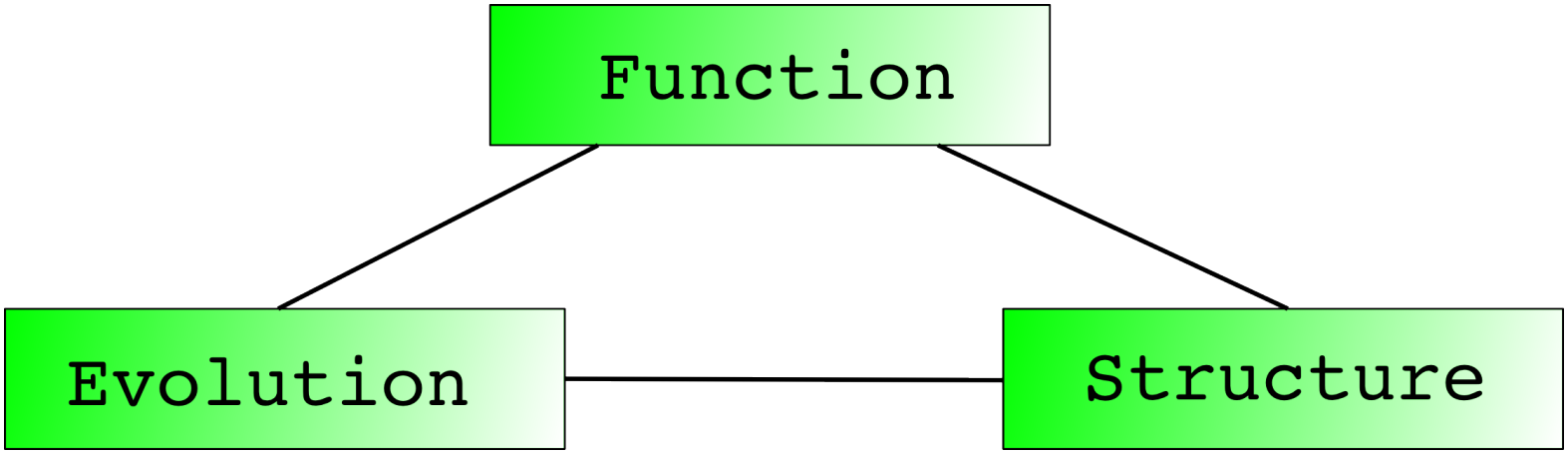
# Introduction

## Which molecules are we interested in?

- *Evolving molecules*: DNA, RNA, Proteins
- Several structures are resolved: we have the set of coordinates for almost all atom positions
- There are several levels of structure organization

| | DNA | RNA | Proteins |
|---|---|---|---|
| I | desoxyribonucleotides sequence | ribonucleotides sequence | amino-acids sequence |
| II | double-helix | loop and stem (double stranded regions) | loop, helices, turns, strands and sheets |
| III | — | yes, for rRNA and tRNA | domain organization |
| IV | — | rRNA | a lot of protein complexes |

# Why should we study their structure?




- The function of a molecule is tightly linked to its structure
- Improve models of sequence evolution, for better inference of evolutionary processes (including phylogeny)
- Use evolutionary information to study/predict the structure of molecules

# Outline of the lecture

1. On the non-homogeneity of the substitution process
   - Rate across site variation
   - Rate across site co-variation
   - Accounting for secondary structure
2. On the non-independence of substitution events
   - The success story of RNA structure prediction
   - Accounting for phylogeny: substitution mapping
   - Detecting coevolution in proteins
   - Non-independence and models of evolution

# All positions are not equal

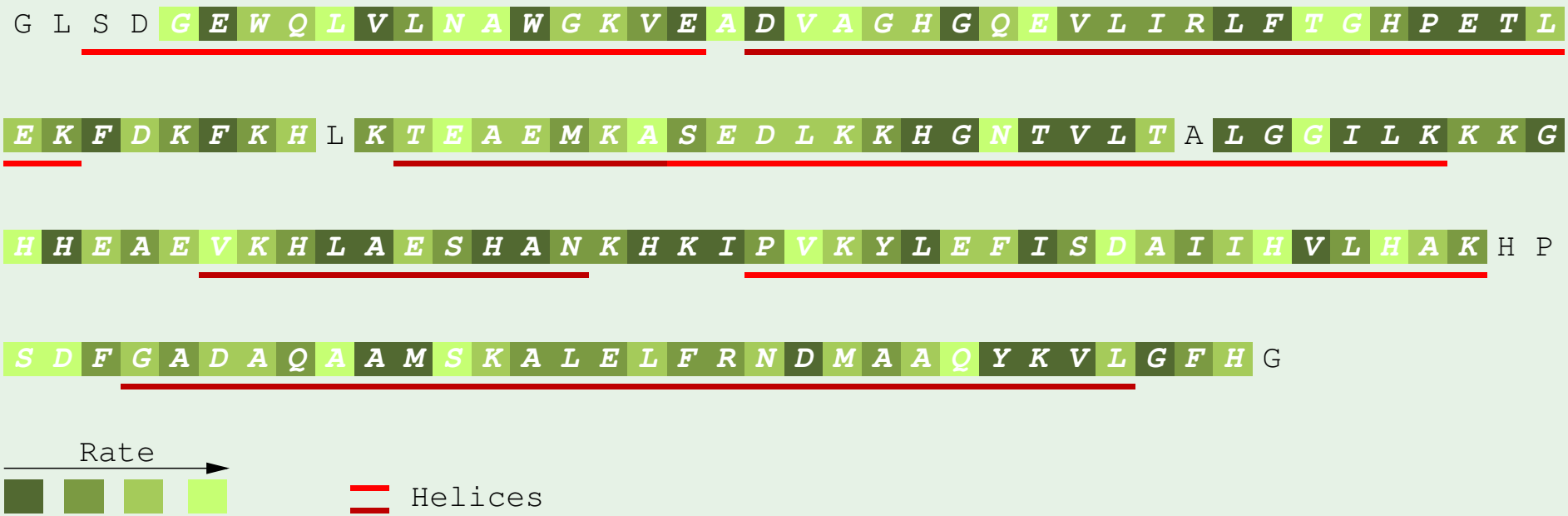
- The structure of a molecule determines a set of constraints acting on individual sites
- The neutral theory states that the level of constraint determines the rate of substitution
- $\Rightarrow$ All positions do not evolve at the same rate

## Estimating site-specific rates

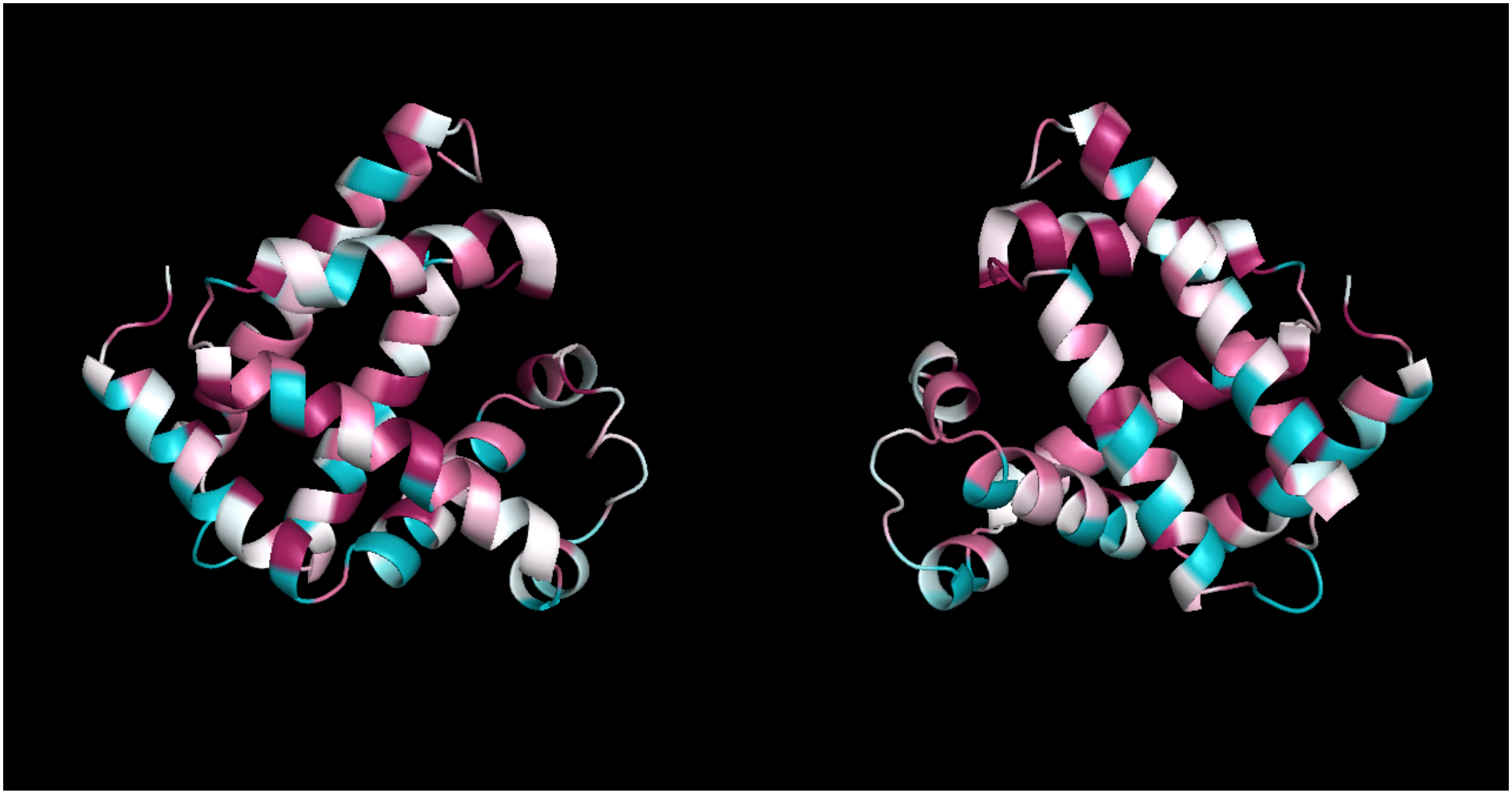
- Site variability (entropy, information): do not account for phylogeny! May be very inaccurate...
- Parsimony score
- Empirical Bayesian estimation (work of Ziheng Yang)

# All positions are not equal

## Example (A vertebrate myoglobin data set)



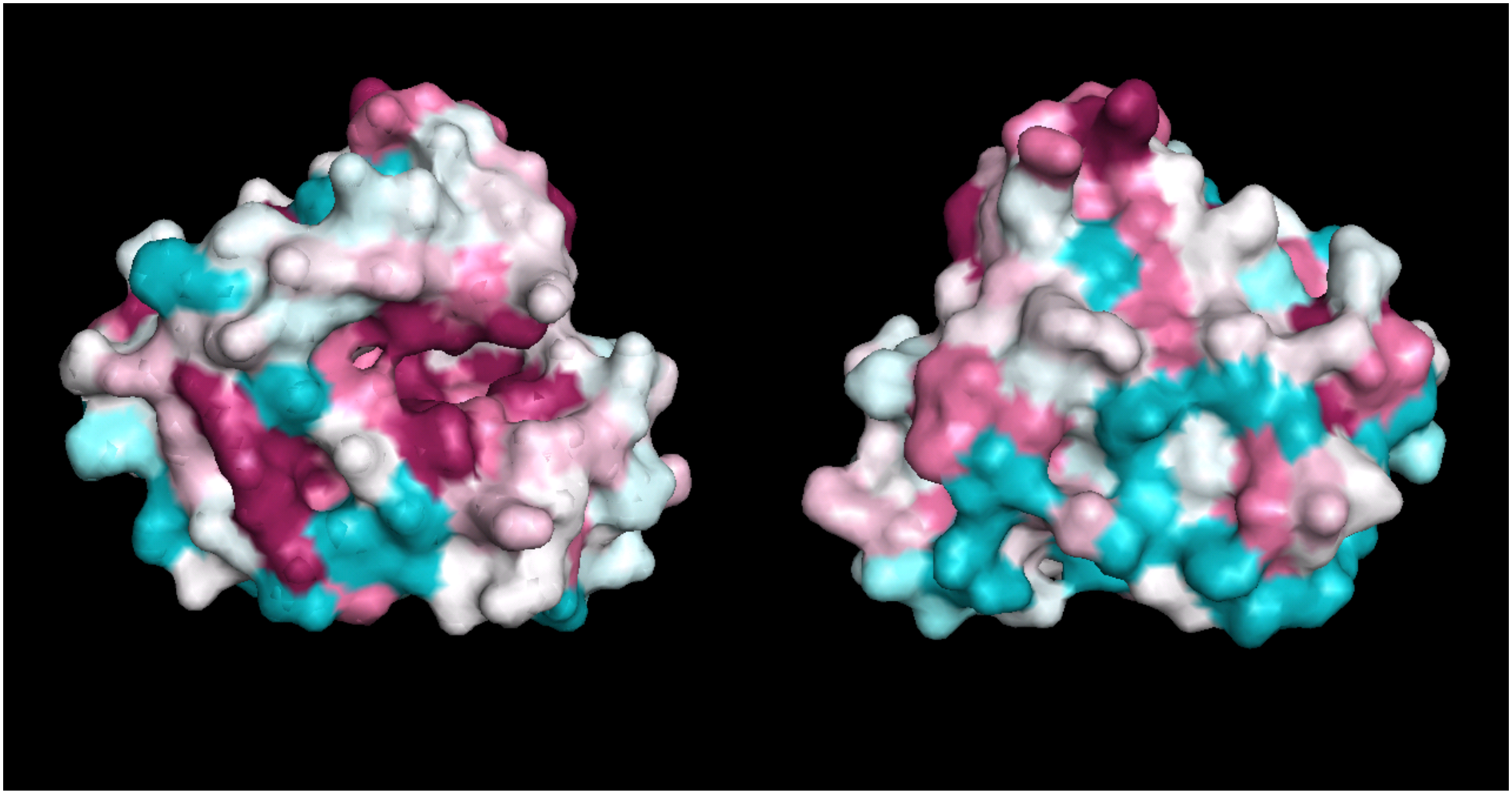
# All positions are not equal



Purple (lent) $\longrightarrow$ Cyan (fast)

Made with ConSurf `http://consurf.tau.ac.il/` [Glaser et al., 2003]

# All positions are not equal



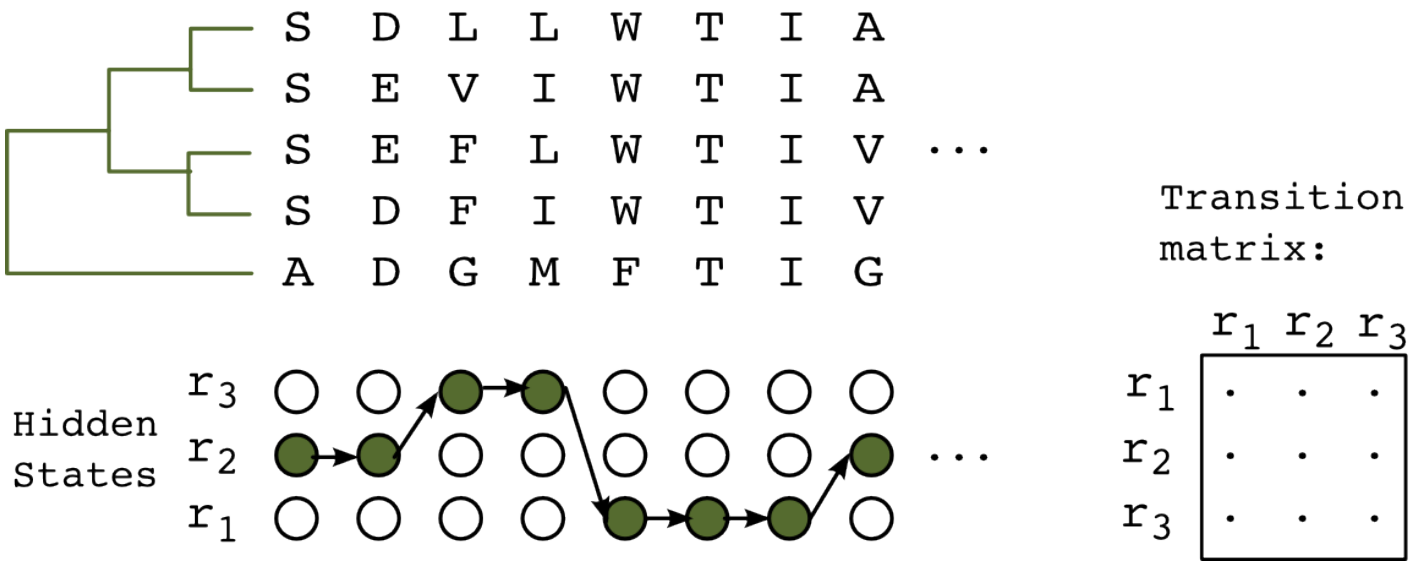
Purple (lent) $\longrightarrow$ Cyan (fast)

Made with ConSurf `http://consurf.tau.ac.il/` [Glaser et al., 2003]

# Do constraints act on individual sites?

- Several motifs cover whole regions of the sequence (helices, strands)
- We then expect the rates of substitutions to be somehow correlated along the sequence
- Model which allow consecutive sites to be correlated: Hidden Markov Models

# Hidden Markov Models: Felsenstein and Churchill [1996]




- The rate of a specific site is unknown, be we can compute the likelihood of a given site according to a given rate
- Each rate is a (hidden) state, and we have a (Markov) model of transitions between states along the alignment
- We can compute the likelihood of the whole data set according to the transition model, estimate transition parameters, estimate the most likely hidden state for each site, etc

# Do constraints act on individual sites?

- Several motifs cover whole regions of the sequence (helices, strands)
- We then expect the rates of substitutions to be somehow correlated along the sequence
- Model which allow consecutive sites to be correlated: Hidden Markov Models

## Hidden Markov Models: Felsenstein and Churchill [1996]'s model

- Several rate classes like in Yang's model
- Model of substitution between rates
- Developed for DNA sequences, application to hemoglobin: helices evolve more rapidly than coils, but large variations

# Distinct structural constraints, distinct substitution processes

## Goldman, Thorne and Jones's model [Thorne et al., 1996, Goldman et al., 1996, 1998]

- Four types of secondary structure: Helix, Sheet, Turn and Coil
- Position dependent states $(1,2,3,\ldots,n-2,n-1,n)$
- Two solvent accessibility classes: Exposed or Buried $\Rightarrow 38 \neq$ states

# Distinct structural constraints, distinct substitution processes

## Goldman, Thorne and Jones's model [Thorne et al., 1996, Goldman et al., 1996, 1998]

- Four types of secondary structure: Helix, Sheet, Turn and Coil
- Position dependent states (1,2,3,...,$n-2$,$n-1$,$n$)
- Two solvent accessibility classes: Exposed or Buried $\Rightarrow$ 38 $\neq$ states

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

- A secondary structure specific replacement matrix is estimated from a set of protein data set with known structure
- This model provides a significantly better fit than independent, homogeneous models

# Distinct structural constraints, distinct substitution processes

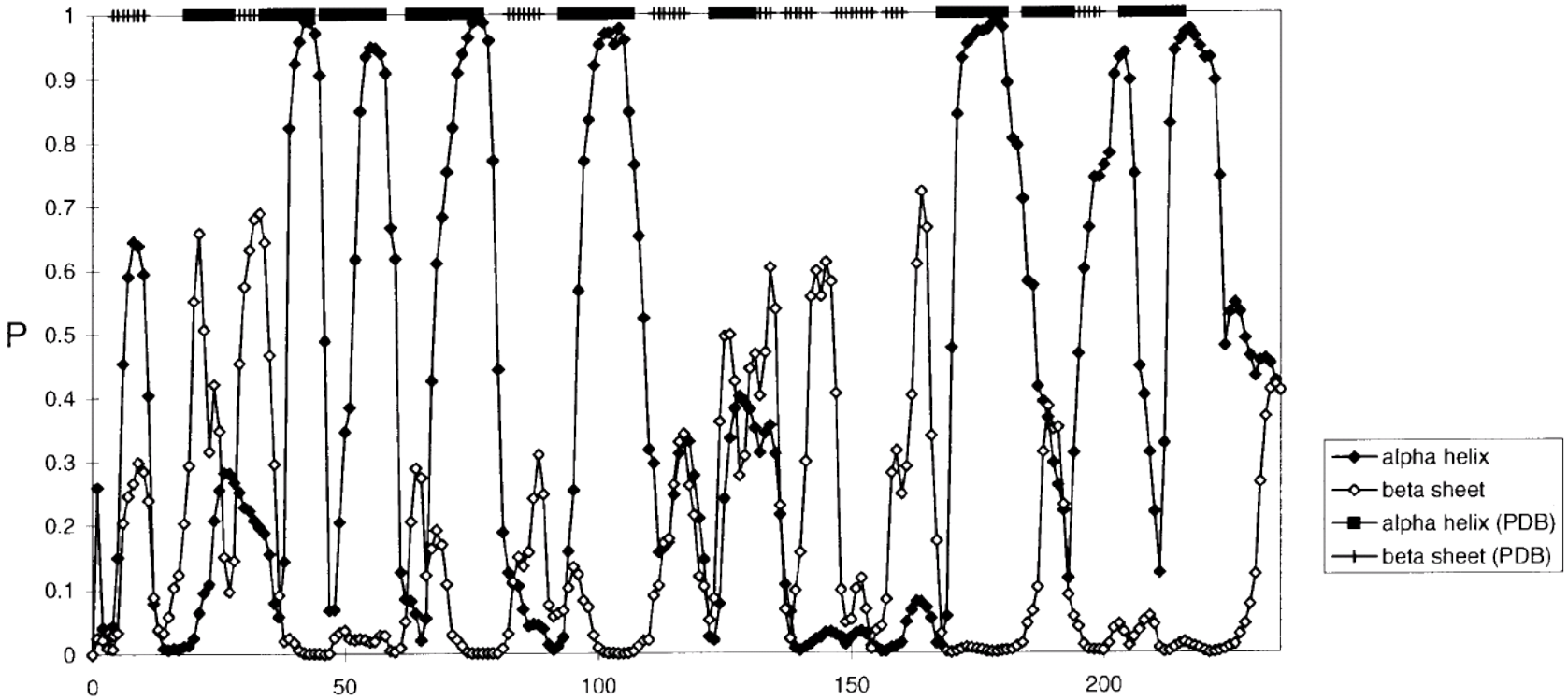
## Goldman, Thorne and Jones's model [Thorne et al., 1996, Goldman et al., 1996, 1998]

- Four types of secondary structure: Helix, Sheet, Turn and Coil
- Position dependent states (1,2,3,...,$n-2$,$n-1$,$n$)
- Two solvent accessibility classes: Exposed or Buried $\Rightarrow$ 38 $\neq$ states

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

- A secondary structure specific replacement matrix is estimated from a set of protein data set with known structure
- This model provides a significantly better fit than independent, homogeneous models
- These matrices can be used with a protein with unknown structure, and allow to predict the most likely motif for each residue

# Example of posterior decoding:

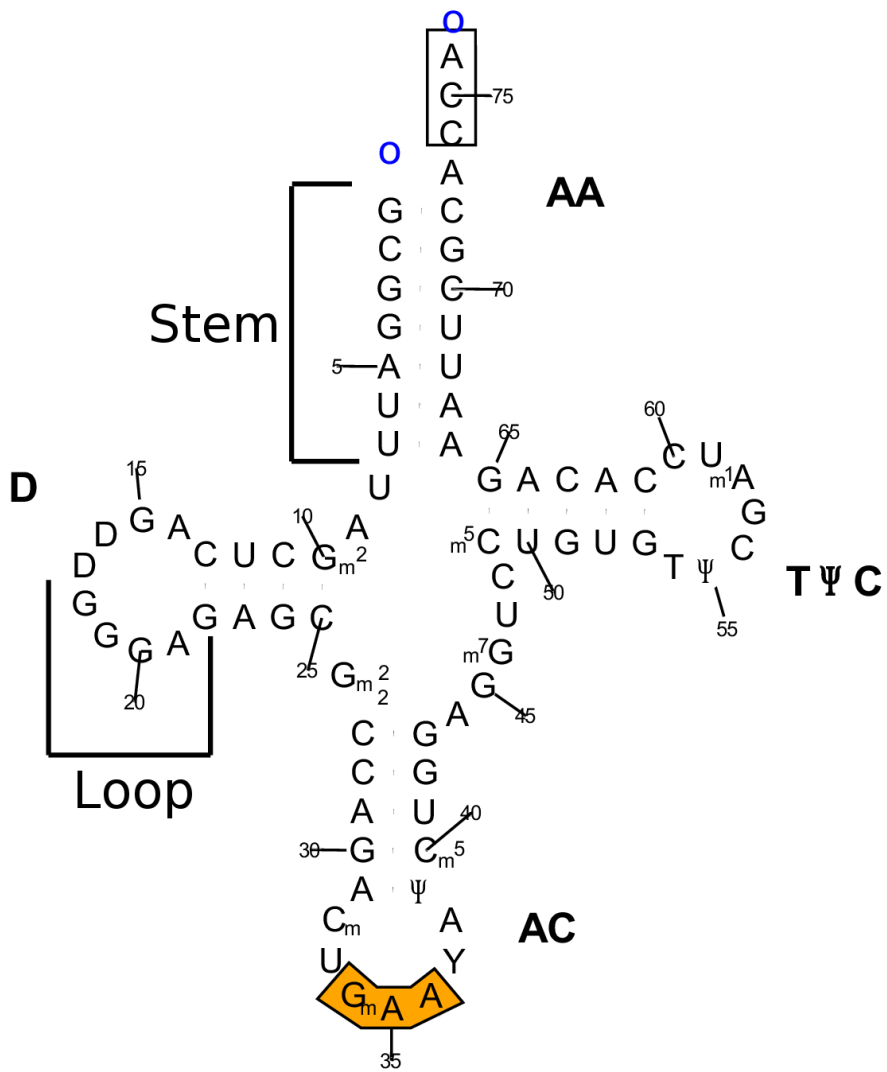


# RNA structure

- Two motifs in RNA: stems and loops
- Constraints on Watson-Crick pairs within stems
- $\Rightarrow$ Find the sequence folding that maximizes the number of W-C pairs ( 80s)

Improvements:

- Use a sequence alignment ( 90s)
- Take into account the underlying phylogeny ( 90s +)
- Use complex models that account for spatial dependency (2000)

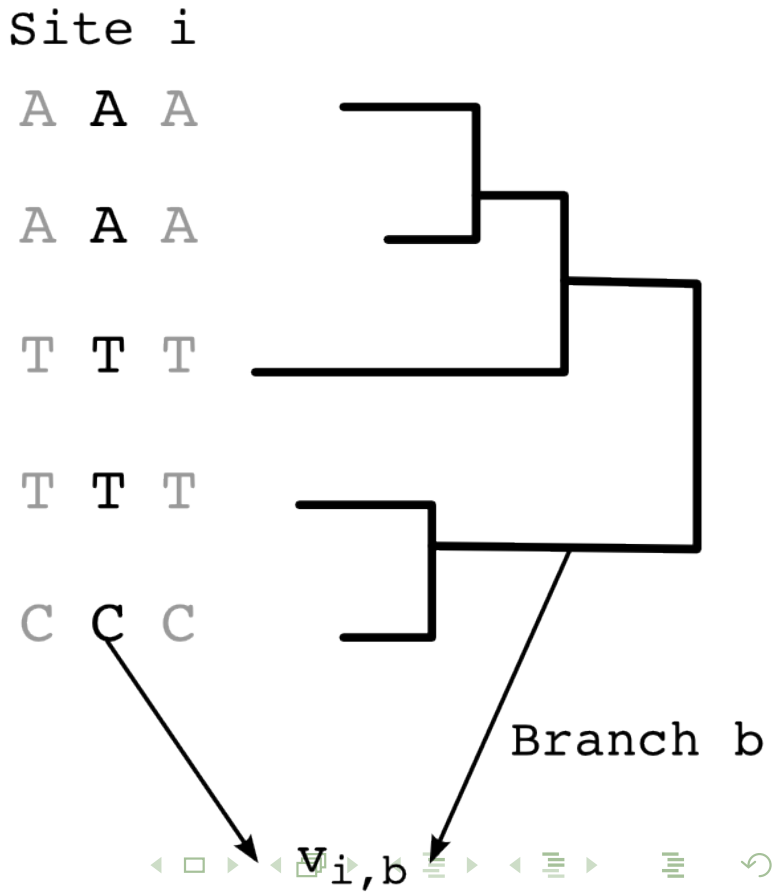


# Using phylogenetics to detect coevolving pairs

## Mapping the substitution events

**Goal:**

Locate the substitution events on a phylogenetic tree for each position of a sequence alignment



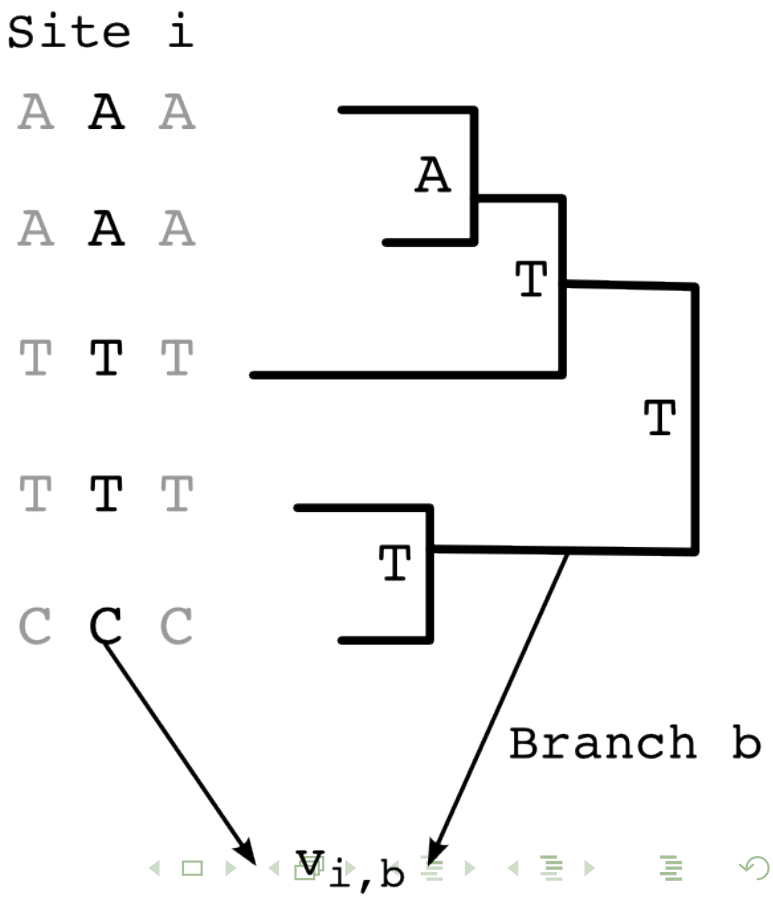
# Using phylogenetics to detect coevolving pairs

## Mapping the substitution events

**Goal:**

Locate the substitution events on a phylogenetic tree for each position of a sequence alignment

- Reconstruct all ancestral states for each node



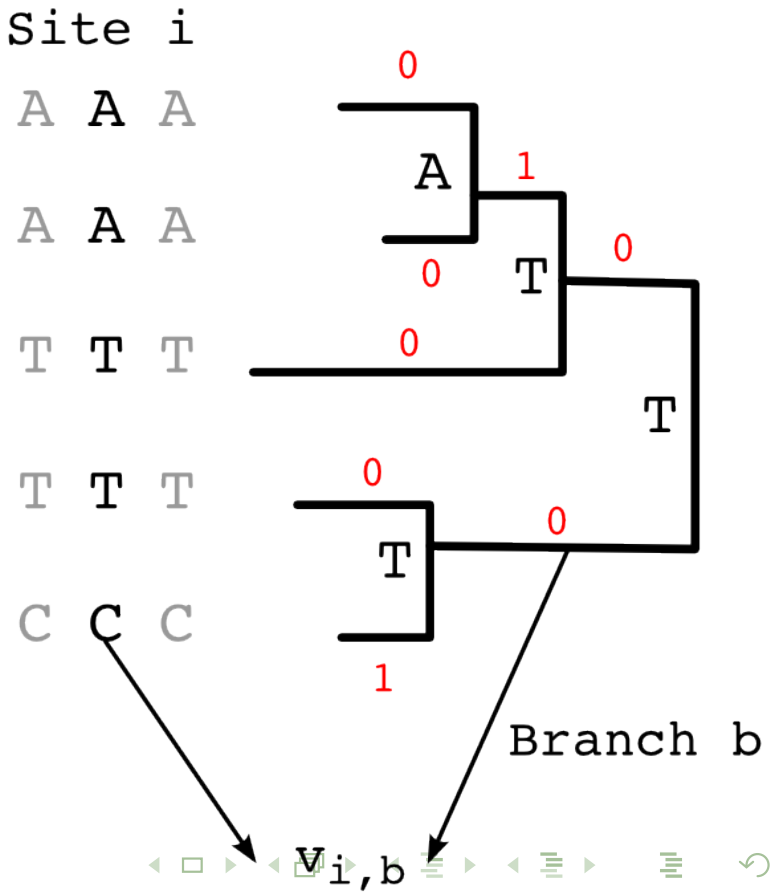
# Using phylogenetics to detect coevolving pairs

## Mapping the substitution events

**Goal:**

Locate the substitution events on a phylogenetic tree for each position of a sequence alignment

- Reconstruct all ancestral states for each node
- For each branch and for each site, count 1 substitution if the states are different, 0 otherwise



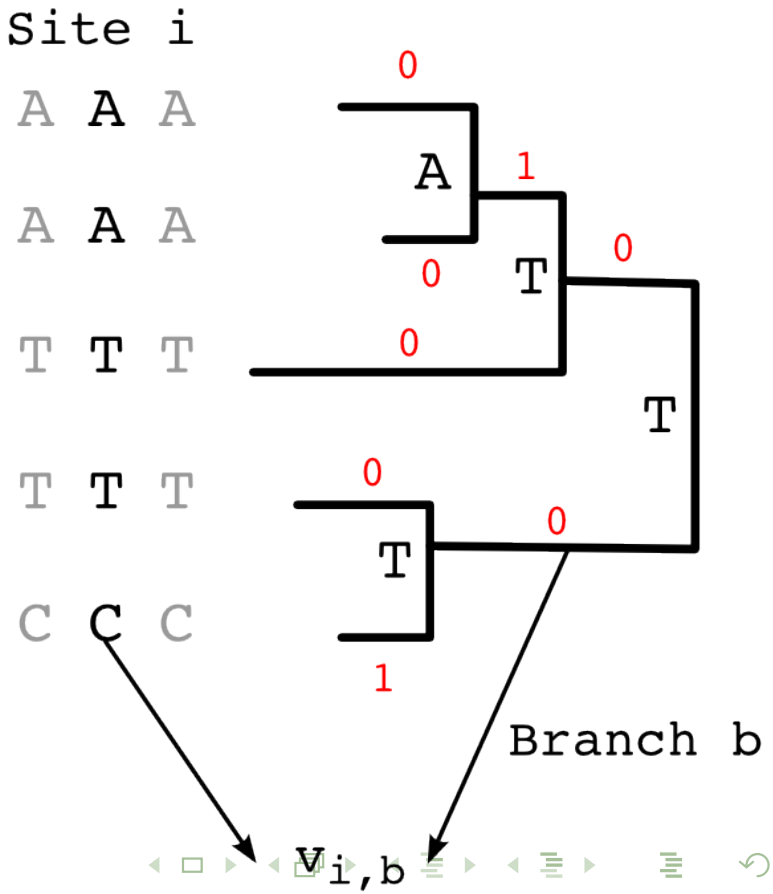
# Using phylogenetics to detect coevolving pairs

## Mapping the substitution events

**Goal:**

Locate the substitution events on a phylogenetic tree for each position of a sequence alignment

- Reconstruct all ancestral states for each node
- For each branch and for each site, count 1 substitution if the states are different, 0 otherwise
- Improvements: take into account uncertainty on ancestral states and branch lengths



# Detecting non-independent positions

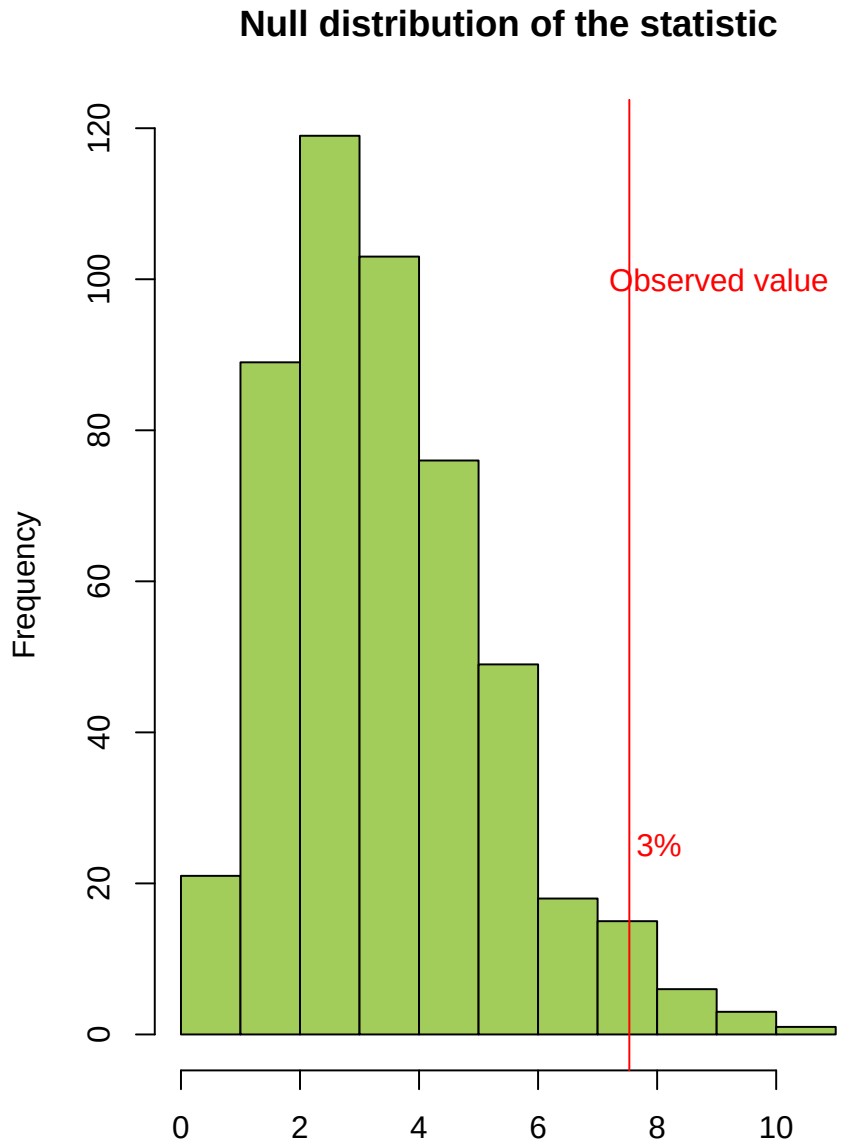
1. Define a measure of coevolution for a group of sites, based on the underlying substitution mapping

# Detecting non-independent positions

1. Define a measure of coevolution for a group of sites, based on the underlying substitution mapping
2. Repeat 1000 times:
   1. Simulate a group of sites under the hypothesis of independence
   2. Record the coevolution measure of the group obtained

# Detecting non-independent positions

1. Define a measure of coevolution for a group of sites, based on the underlying substitution mapping
2. Repeat 1000 times:
   1. Simulate a group of sites under the hypothesis of independence
   2. Record the coevolution measure of the group obtained
3. Compare the value of the coevolution for real groups and the ones obtained from simulations

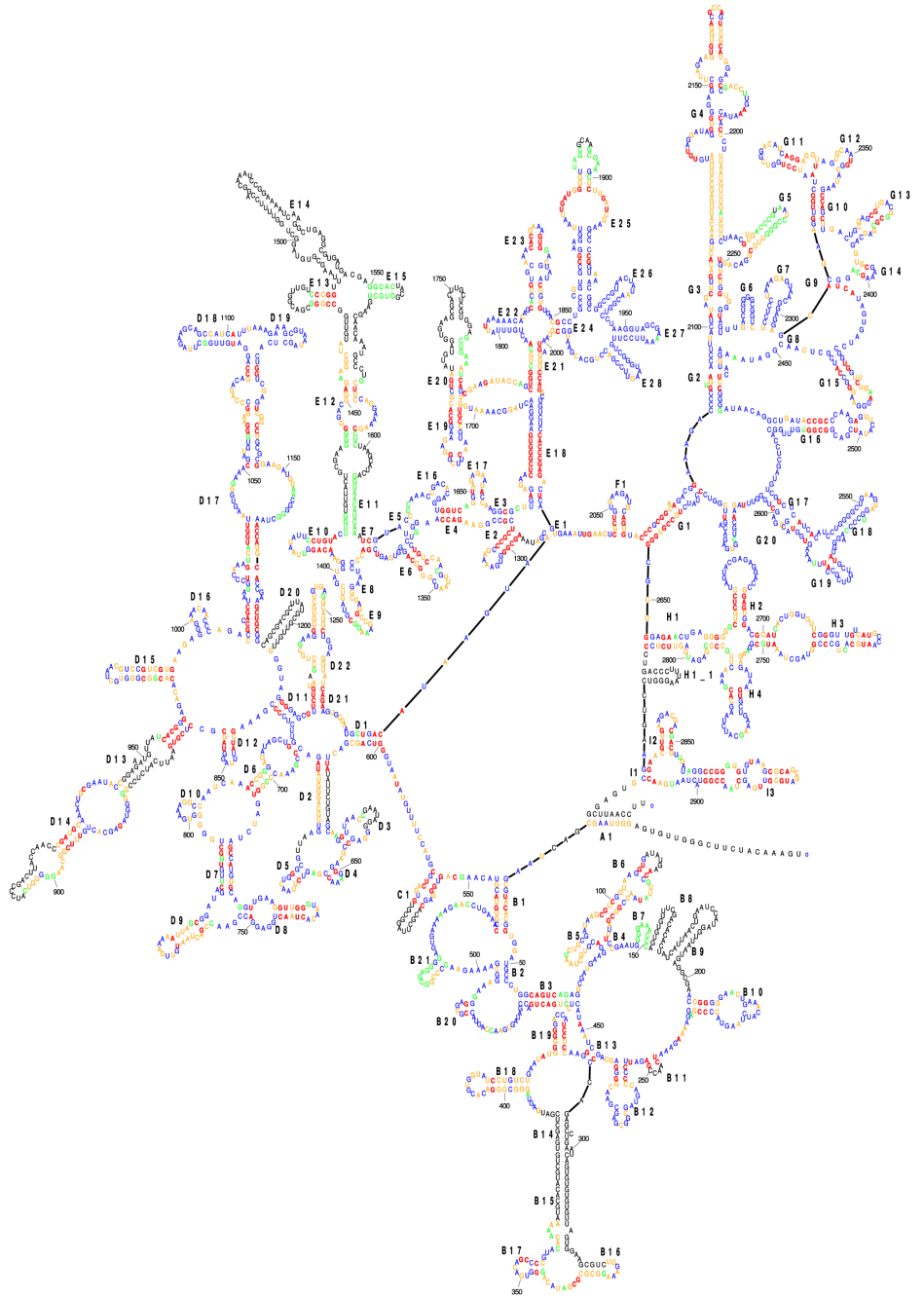


# A Bacteria example

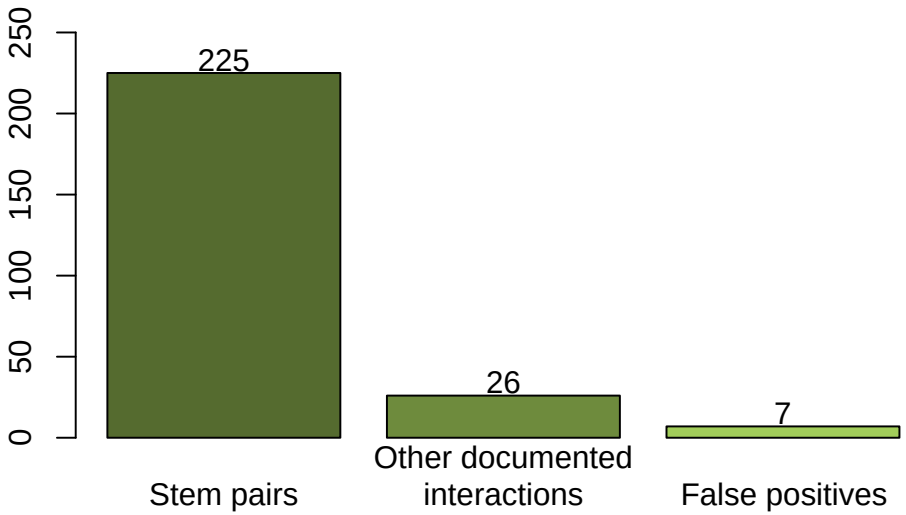
- 80 sequences of LSU, all possible pairs tested

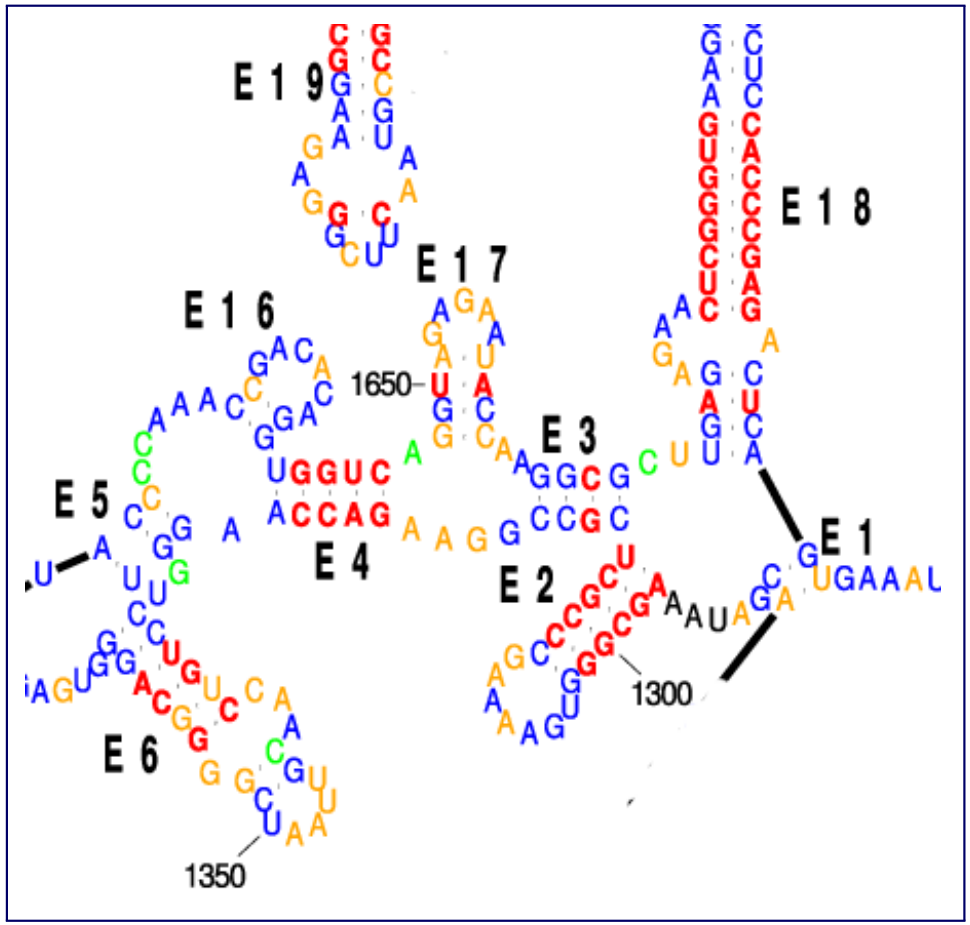
# A Bacteria example

- 80 sequences of LSU, all possible pairs tested
- 258 pairs show significant coevolution,



# A Bacteria example




- 80 sequences of LSU, all possible pairs tested
- 258 pairs show significant coevolution,
- 225 belong to secondary structure, 26 to tertiary structure

# The protein case

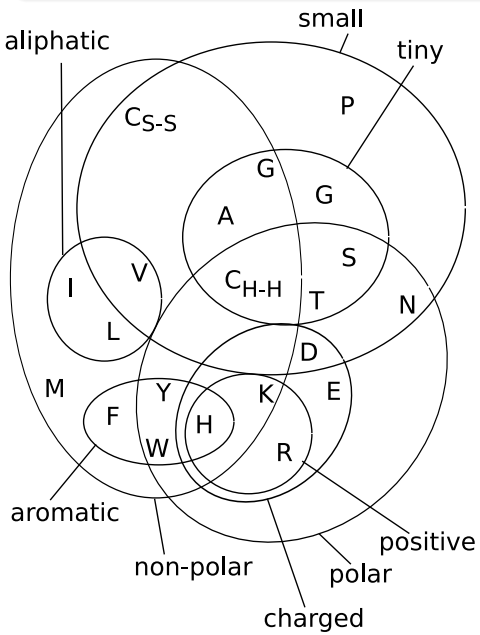
## More challenging:

- 20 possible states instead of 4

# The protein case

## More challenging:

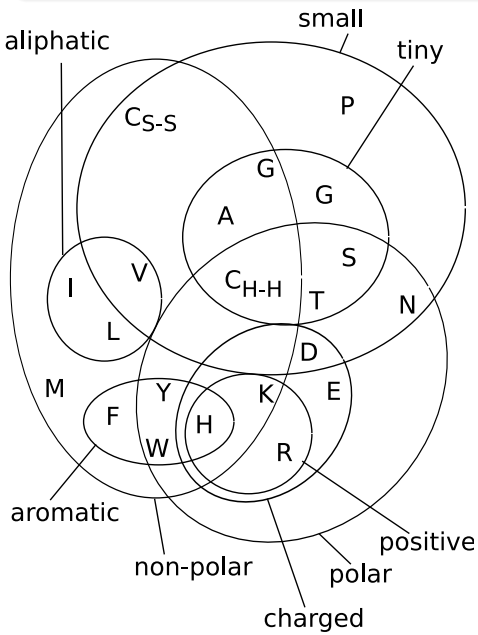
- 20 possible states instead of 4
- Several biochemical properties to compensate for

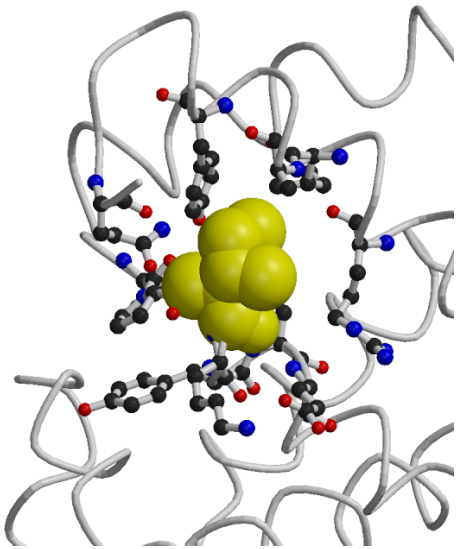


# The protein case

## More challenging:

- 20 possible states instead of 4
- Several biochemical properties to compensate for
- Larger connectivity of residues

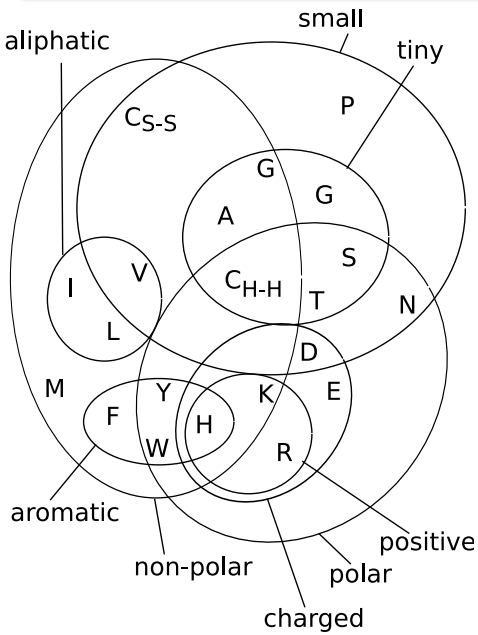


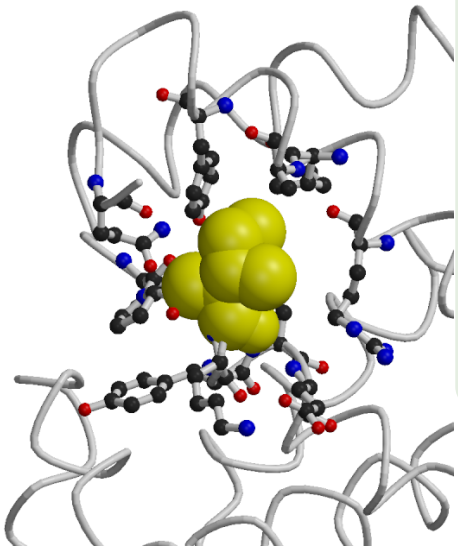



# The protein case

## More challenging:

- 20 possible states instead of 4
- Several biochemical properties to compensate for
- Larger connectivity of residues



## Main results:

- Signal is scarcer, needs more data to detect
- Positions in close proximity tend to coevolve more than distant positions
- Relation with secondary structure is unclear
- Coevolution within domains is higher than between domains

# Non-independence and models of evolution

## Phylogenetic inference:

- ML is robust to departures from independence, NJ more sensitive:

**Neighbor Joining and Maximum Likelihood with RNA Sequences: Addressing the Interdependence of Sites**

*Elisabeth R. M. Tillier* and Richard A. Collins*

*Department of Botany and Department of Molecular and Medical Genetics, University of Toronto and Canadian Institute for Advanced Research Program in Evolutionary Biology

- Inter-dependence artificially increases the bootstrap support:

*Syst. Biol.* 53(1):38–46, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490264680

**Sampling Properties of the Bootstrap Support in Molecular Phylogeny: Influence of Nonindependence Among Sites**

NICOLAS GALTIER

*CNRS UMR 5000, Génome, Populations, Interactions, Université Montpellier 2, CC 63, Place E. Bataillon, 34095 Montpellier, France; E-mail: galtier@univ-montp2.fr*

# A simple model of non-independent evolution: RNA

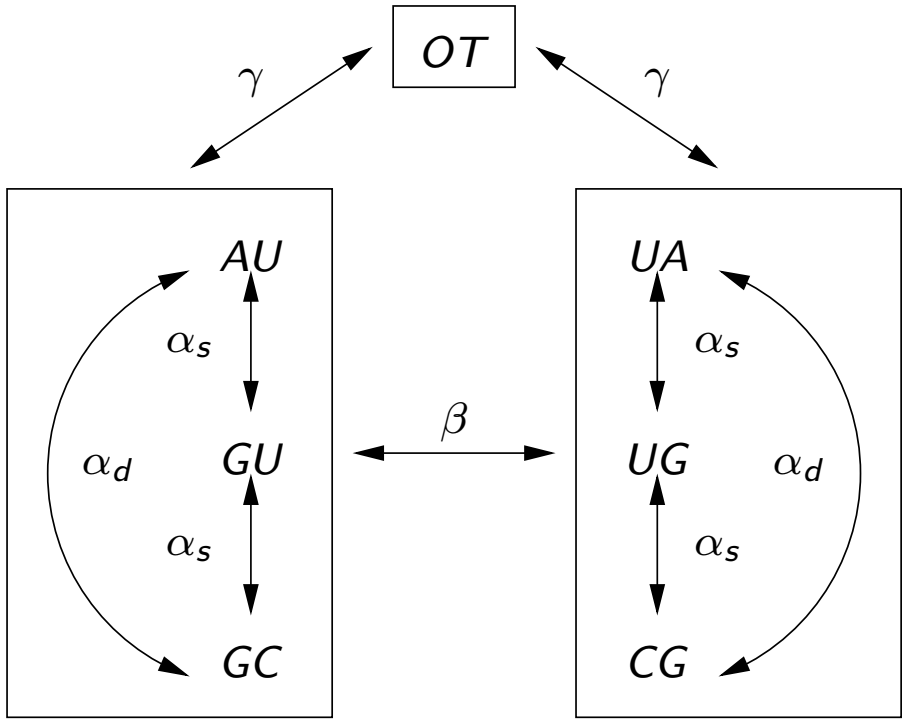
Tillier and Collins [1998]

- Model with pairs of states:
  $4 \times 4 = 16$ states

# A simple model of non-independent evolution: RNA

Tillier and Collins [1998]

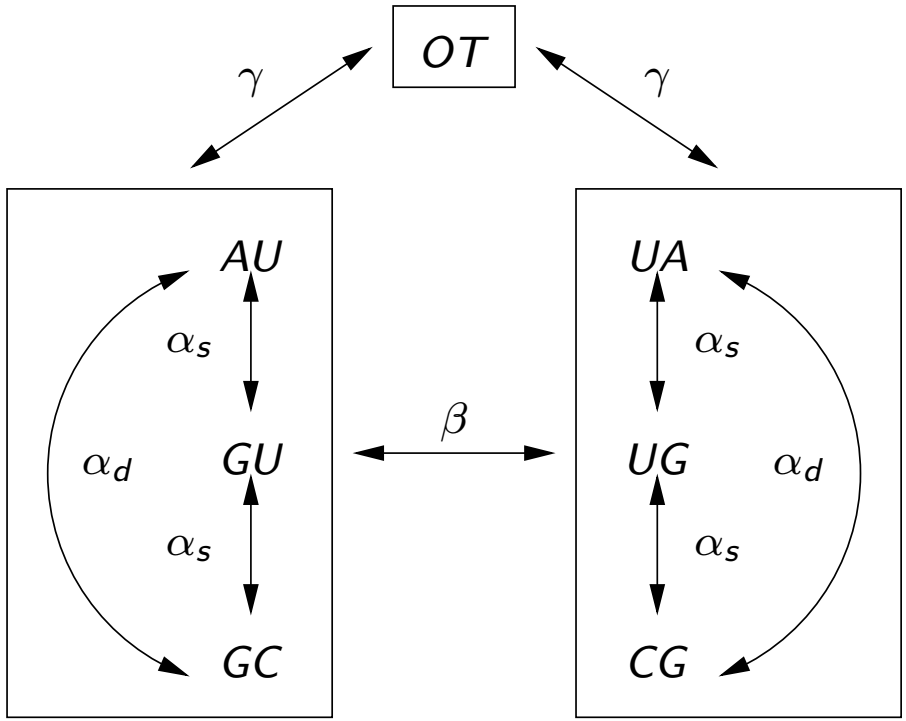
- Model with pairs of states: $4 \times 4 = 16$ states
- Model simplification, Watson-Crick pairs and GU intermediates



# A simple model of non-independent evolution: RNA

Tillier and Collins [1998]

- Model with pairs of states: $4 \times 4 = 16$ states
- Model simplification, Watson-Crick pairs and GU intermediates
- Results: high substitution rate, compensating mutations appear almost simultaneously



# A simple model of non-independent evolution: proteins

Pollock et al. [1999]

- Model with pairs of states: $20 \times 20 =$ two many pairs of state!

# A simple model of non-independent evolution: proteins

Pollock et al. [1999]

- Model with pairs of states: $20 \times 20 =$ two many pairs of state!
- Model simplification, using sub-alphabets, for instance:
  - Large/Small
  - Polar/Non-polar
  - Charged/Non-charged

$$Q = \begin{array}{c} \\ AB \\ Ab \\ aB \\ ab \end{array} \begin{array}{c} \begin{array}{cccc} AB & Ab & aB & ab \end{array} \\ \begin{pmatrix} -\sum_{AB} & \lambda_B \pi_{Ab}/\pi_A & \lambda_A \pi_{aB}/\pi_B & 0 \\ \lambda_B \pi_{AB}/\pi_A & -\sum_{Ab} & 0 & \lambda_A \pi_{ab}/\pi_b \\ \lambda_A \pi_{AB}/\pi_B & 0 & -\sum_{aB} & \lambda_B \pi_{ab}/\pi_a \\ 0 & \lambda_A \pi_{ab}/\pi_b & \lambda_B \pi_{aB}/\pi_a & -\sum_{ab} \end{pmatrix} \end{array}$$

# A simple model of non-independent evolution: proteins

Pollock et al. [1999]

- Model with pairs of states: $20 \times 20 =$ two many pairs of state!
- Model simplification, using sub-alphabets, for instance:
  - Large/Small
  - Polar/Non-polar
  - Charged/Non-charged

$$Q = \begin{array}{c} \\ AB \\ Ab \\ aB \\ ab \end{array} \begin{array}{c} \begin{array}{cccc} AB & Ab & aB & ab \end{array} \\ \begin{pmatrix} -\sum_{AB} & \lambda_B \pi_{Ab}/\pi_A & \lambda_A \pi_{aB}/\pi_B & 0 \\ \lambda_B \pi_{AB}/\pi_A & -\sum_{Ab} & 0 & \lambda_A \pi_{ab}/\pi_b \\ \lambda_A \pi_{AB}/\pi_B & 0 & -\sum_{aB} & \lambda_B \pi_{ab}/\pi_a \\ 0 & \lambda_A \pi_{ab}/\pi_b & \lambda_B \pi_{aB}/\pi_a & -\sum_{ab} \end{pmatrix} \end{array}$$

- Computationally demanding: used only to detect significant coevolving pairs by model comparison.

# A more complex model of non-independent evolution

Rodrigue et al. [2005, 2006]

- Sites are treated simultaneously: sequence of size $N \Rightarrow 20^N$ states!

$$R_{xy} = \begin{cases} 0 & \text{if sequence } x \text{ and } y \text{ differ} \\ & \text{by more than one position} \\ Q_{kl} & \text{if sequence } x \text{ and } y \text{ have states} \\ & l \text{ and } m \text{ at positions } i \text{ and } j \\ -\sum_{y \neq x} R_{xy} & \text{i} = \text{j} \end{cases}$$

# A more complex model of non-independent evolution

Rodrigue et al. [2005, 2006]

- Sites are treated simultaneously: sequence of size $N \Rightarrow 20^N$ states!

$$R_{xy} = \begin{cases} 0 & \text{if sequence } x \text{ and } y \text{ differ by more than one position} \\ Q_{kl} e^{p \times (E(x) - E(y))} & \text{if sequence } x \text{ and } y \text{ have states } l \text{ and } m \text{ at positions } i \text{ and } j \\ -\sum_{y \neq x} R_{xy} & \text{i = j} \end{cases}$$

- Add a fitness function that takes into account the whole sequence. This function is computed using measures on real data: statistical potentials. $p$ is a free parameter, estimated from the data.

# A more complex model of non-independent evolution

Rodrigue et al. [2005, 2006]

- Sites are treated simultaneously: sequence of size $N \Rightarrow 20^N$ states!

$$R_{xy} = \begin{cases} 0 & \text{if sequence } x \text{ and } y \text{ differ} \\ & \text{by more than one position} \\ Q_{kl} e^{p \times (E(x) - E(y))} & \text{if sequence } x \text{ and } y \text{ have states} \\ & l \text{ and } m \text{ at positions } i \text{ and } j \\ -\sum_{y \neq x} R_{xy} & \text{i = j} \end{cases}$$

- Add a fitness function that takes into account the whole sequence. This function is computed using measures on real data: statistical potentials. $p$ is a free parameter, estimated from the data.
- Bayesian sampling procedure to estimate parameters.

# A more complex model of non-independent evolution

Rodrigue et al. [2005, 2006]

- Sites are treated simultaneously: sequence of size $N \Rightarrow 20^N$ states!

$$R_{xy} = \begin{cases} 0 & \text{if sequence } x \text{ and } y \text{ differ by more than one position} \\ Q_{kl} e^{p \times (E(x) - E(y))} & \text{if sequence } x \text{ and } y \text{ have states } l \text{ and } m \text{ at positions } i \text{ and } j \\ -\sum_{y \neq x} R_{xy} & \text{i = j} \end{cases}$$

- Add a fitness function that takes into account the whole sequence. This function is computed using measures on real data: statistical potentials. $p$ is a free parameter, estimated from the data.
- Bayesian sampling procedure to estimate parameters.
- Usage limited to small data sets due to the computational load, but provides a better description of the data.

# Conclusions

- Large effort during the last 10 years
- Methodological improvements are limited by the large complexity of structural constraints
- The large amount of sequence and structure data opens a way to characterize these constraints at a large scale, and maybe help the design a new tractable models
- These studies are also a good example of tight interactions between evolutionary biology, molecular biology and bioinformatics.

# References

J. Felsenstein and G. A. Churchill. A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13:93–104, 1996.

F. Glaser, T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, 19:163–164, 2003.

N. Goldman, J. L. Thorne, and D. T. Jones. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *Journal of Molecular Biology*, 263:196–208, 1996.

N. Goldman, J. L. Thorne, and D. T. Jones. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149:445–458, 1998.

D. D. Pollock, W. R. Taylor, and N. Goldman. Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology*, 287:187–198, 1999.

N. Rodrigue, N. Lartillot, D. Bryant, and H. Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347:207–217, 2005.

N. Rodrigue, H. Philippe, and N. Lartillot. Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular Biology and Evolution*, 23:1762–1775, 2006.

J. L. Thorne, N. Goldman, and D. T. Jones. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13:666–673, 1996.

E. R. M. Tillier and R. A. Collins. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal rna. *Genetics*, 148:1993–2002, 1998.