AUTOMATED REASONING

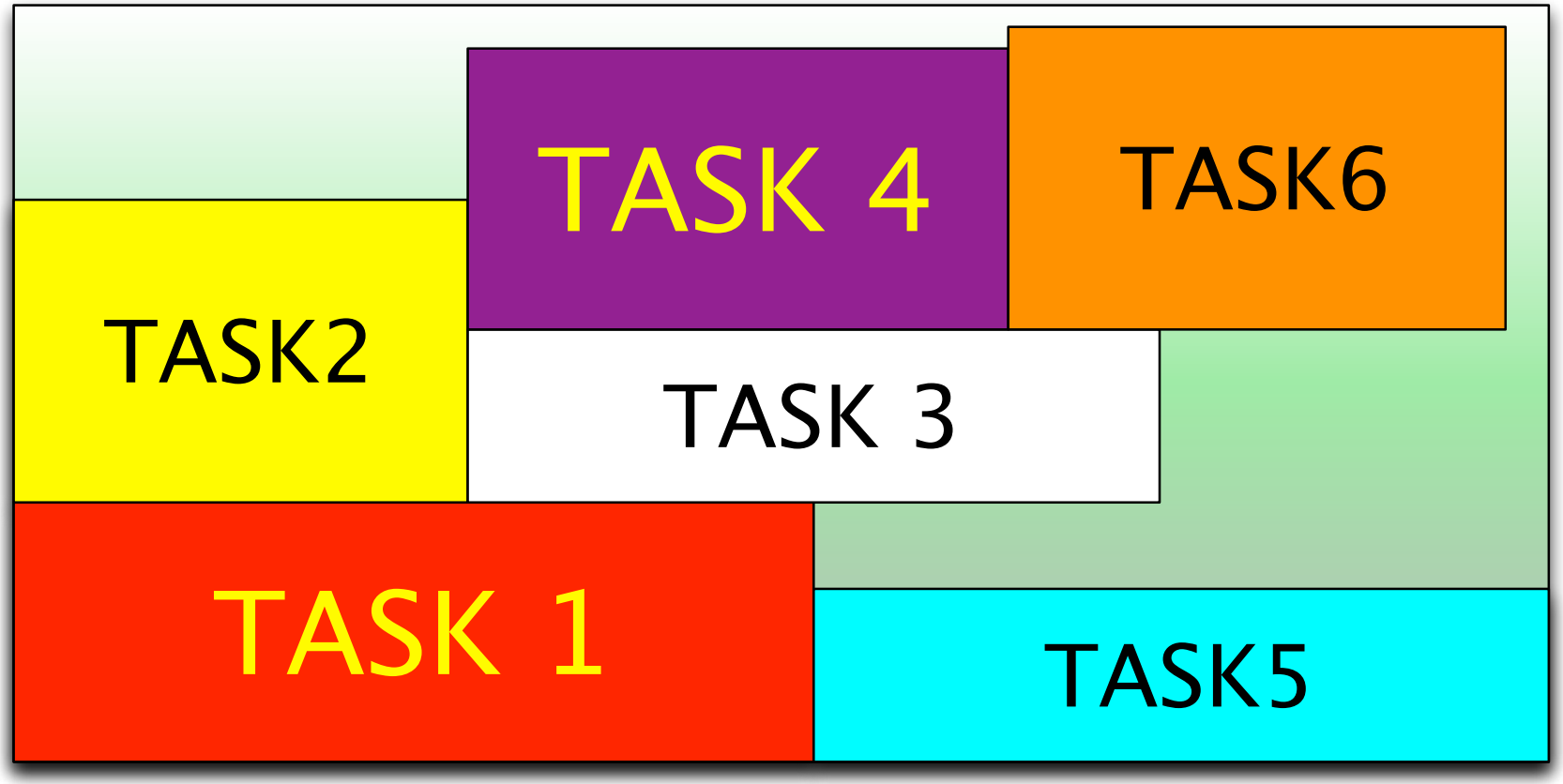
Agostino Dovier

Università di Udine
CLPLAB

Udine, October 2016

Cumulative Scheduling

CUMULATIVE SCHEDULING



There are k tasks, where each task has a fixed duration and a fixed amount of use resource. The goal is to find a schedule that minimizes the completion time for the schedule while not exceeding the capacity c of the resource (time in x-axis, resource in y-axis)

CUMULATIVE SCHEDULING

We introduce the problem with a concrete example. There are 7 tasks where each task has a fixed duration and a fixed amount of use resource:

Task	Duration	Resource
1	16	2
2	6	9
3	13	3
4	7	7
5	5	10
6	18	1
7	4	11

The goal is to find a schedule that minimizes the completion time for the schedule while not exceeding the capacity 13 of the resource.

Of course other constraints can be added on precedences between tasks.

CUMULATIVE SCHEDULING

The resource constraint is succinctly captured by `cumulative` global constraint.

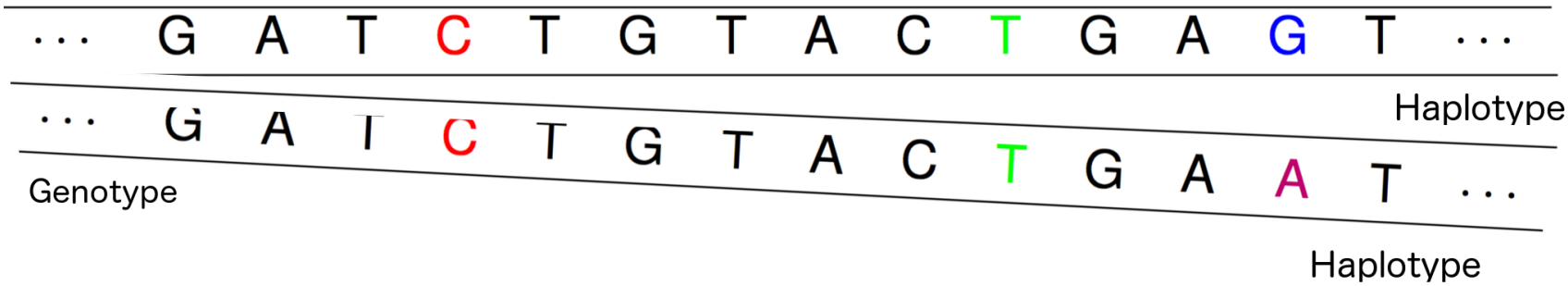
```
cumulative(array[int] of var int: s,
                    array[int] of var int: d,
                    array[int] of var int: r,
                    var int: b)
```

Requires that a set of tasks given by start times s, durations d, and resource requirements r, never require more than a global resource bound b at any one time.
Aborts if s, d, and r do not have identical index sets.
Aborts if a duration or resource requirement is negative.

Haplotype Inference

Haplotype Inference

- Genes are packaged in bundles called chromosomes. (Chromosomes are therefore regions of DNA)
- In Diploid organisms (like humans) there are almost identical chromosome pairs. Each pair is made of an inherited chromosome from the father and another one from the mother.
- A haplotype is a DNA sequence that has been inherited from one parent.
- A genotype is a pairing of two corresponding haplotypes.



HAPLOTYPE INFERENCE

Each person inherits two haplotypes (from the mother and from the father) for most regions of the genome.

$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	G	T	$\cdots$
$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	A	T	$\cdots$

In some typical positions, the bases are subject to mutations.

In the most common case, there is a Single Nucleotide Polymorphism (SNP).

Mutations are $C \leftrightarrow T$ and $A \leftrightarrow G$

HAPLOTYPE INFERENCE

Each person inherits two haplotypes (from the mother and from the father) for most regions of the genome.

$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	G	T	$\cdots$
$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	A	T	$\cdots$

In some typical positions, the bases are subject to mutations.

In the most common case, there is a Single Nucleotide Polymorphism (SNP).

Mutations are $C \leftrightarrow T$ and $A \leftrightarrow G$

HAPLOTYPE INFERENCE

Each person inherits two haplotypes (from the mother and from the father) for most regions of the genome.

$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	G	T	$\cdots$
$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	A	T	$\cdots$

In some typical positions, the bases are subject to mutations.

In the most common case, there is a Single Nucleotide Polymorphism (SNP).

Mutations are $C \leftrightarrow T$ and $A \leftrightarrow G$

Haplotype Inference

- The genome (genotype) is easier and cheaper to be obtained.
- We need computational methods to guess haplotypes
- The Haplotype Inference problem is introduced to investigate genetic variations in a population.
- Typically SNPs sites are the target of the analysis

Haplotype Inference

Single Nucleotide Polymorphism (SNP)

Each person has two haplotypes (from the mother and from the father) for most regions of the genome:

G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	G	T
G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	A	T

Let us focus on the SNPs:

A	C	T	G
A	C	T	A

We encode SNPs according to: $A \mapsto 0$ $C \mapsto 0$ $G \mapsto 1$ $T \mapsto 1$

0	0	1	1
0	0	1	0

0	0	1	2

The genotype is set to 2 if there is a mismatch

Haplotype Inference

Single Nucleotide Polymorphism (SNP)

Each person has two haplotypes (from the mother and from the father) for most regions of the genome:

G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	G	T
G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	A	T

Let us focus on the SNPs:

A	C	T	G
A	C	T	A

We encode SNPs according to: $A \mapsto 0$ $C \mapsto 0$ $G \mapsto 1$ $T \mapsto 1$

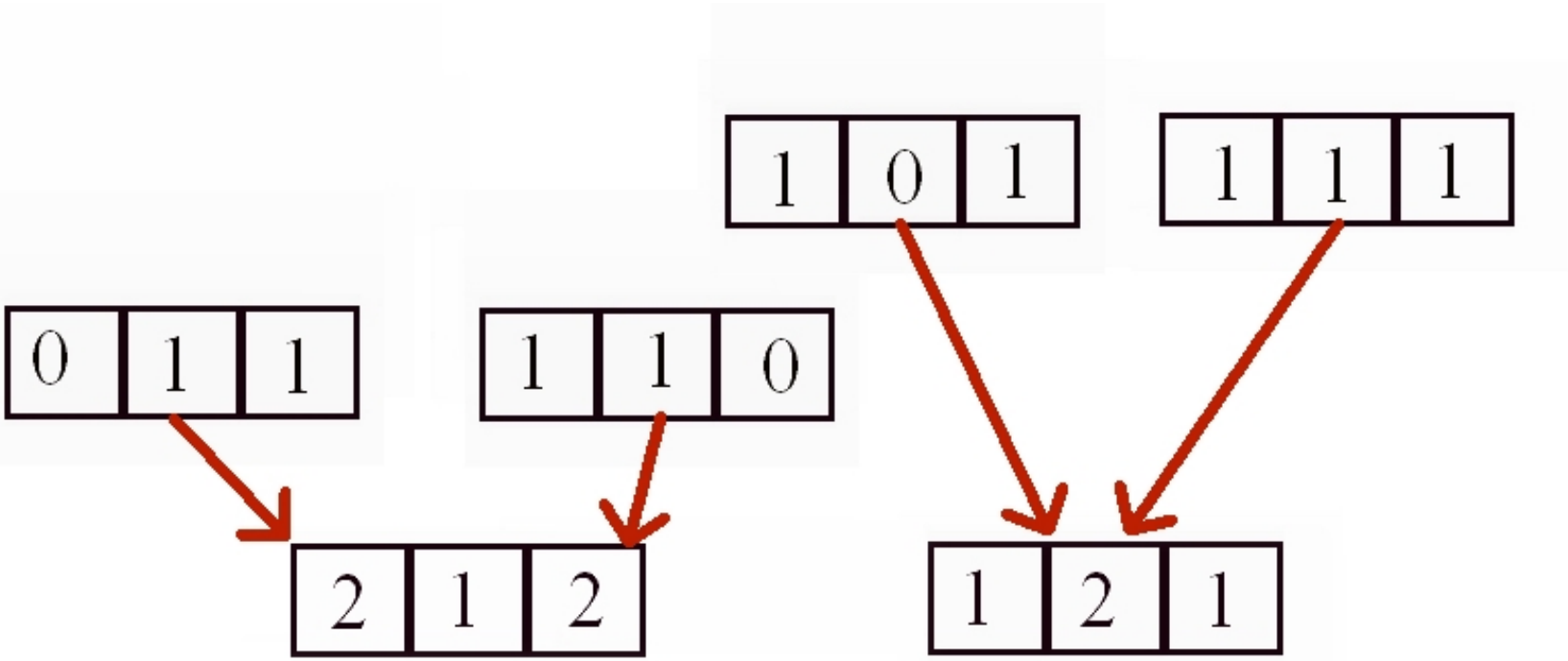
0	0	1	1
0	0	1	0

0	0	1	2

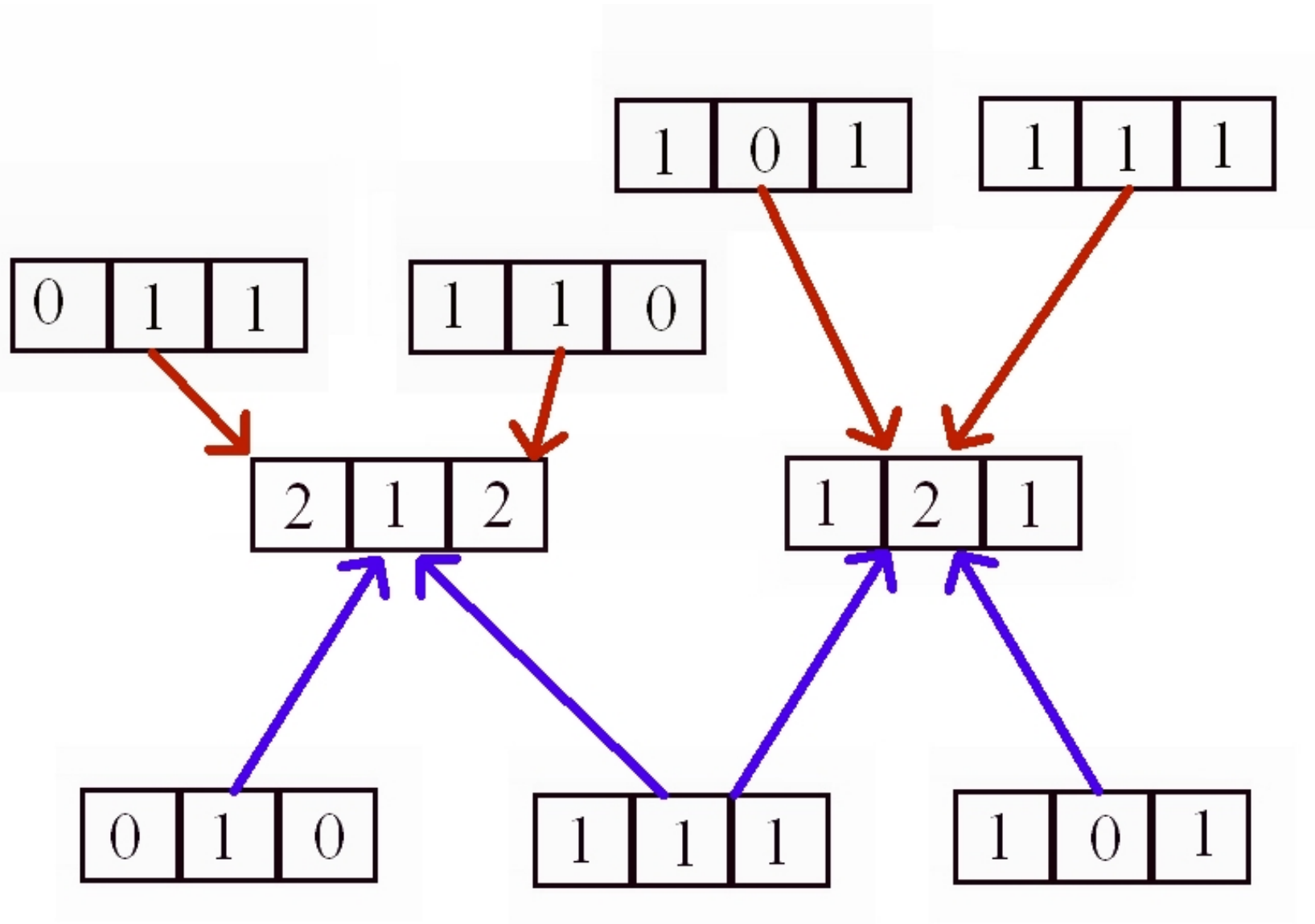
The genotype is set to 2 if there is a mismatch

HAPLOTYPE INFERENCE

HAPLOTYPE INFERENCE



HAPLOTYPE INFERENCE



Haplotype Inference

- A string of $\{0,1\}^*$ is called a *haplotype*
- A string of $\{0,1,2\}^*$ is called a *genotype*
- Two equal length haplotypes generate a unique genotype
- The rules are $0 \oplus 0 = 0$, $1 \oplus 1 = 1$, $0 \oplus 1 = 1 \oplus 0 = 2$
  E.g., $0010, 0101 \Rightarrow 0222$
- If we have a genotype, we can only conjecture (potentially exponentially many) pairs of *haplotypes* that generated it (observe that, e.g., $0110, 0001 \Rightarrow 0222$)
- Biological experiments allow us to know genotypes!
- Since genotypes are introduced in evolution, it is reasonable to find minimal sets of haplotypes explaining the known genotypes.

HAPLOTYPE INFERENCE

- A string of $\{0,1\}^*$ is called a *haplotype*
- A string of $\{0,1,2\}^*$ is called a *genotype*
- Two equal length haplotypes generate a unique genotype
- The rules are $0 \oplus 0 = 0$, $1 \oplus 1 = 1$, $0 \oplus 1 = 1 \oplus 0 = 2$
  E.g., $0010, 0101 \Rightarrow 0222$
- If we have a genotype, we can only conjecture (potentially exponentially many) pairs of *haplotypes* that generated it (observe that, e.g., $0110, 0001 \Rightarrow 0222$)
- Biological experiments allow us to know genotypes!
- Since genotypes are introduced in evolution, it is reasonable to find minimal sets of haplotypes explaining the known genotypes.

HAPLOTYPE INFERENCE

- A string of $\{0,1\}^*$ is called a *haplotype*
- A string of $\{0,1,2\}^*$ is called a *genotype*
- Two equal length haplotypes generate a unique genotype
- The rules are $0 \oplus 0 = 0$, $1 \oplus 1 = 1$, $0 \oplus 1 = 1 \oplus 0 = 2$
  E.g., $0010, 0101 \Rightarrow 0222$
- If we have a genotype, we can only conjecture (potentially exponentially many) pairs of*haplotype*s that generated it (observe that, e.g., $0110, 0001 \Rightarrow 0222$)
- Biological experiments allow us to know genotypes!
- Since genotypes are introduced in evolution, it is reasonable to find minimal sets of haplotypes explaining the known genotypes.

HAPLOTYPE INFERENCE

- A string of $\{0,1\}^*$ is called a *haplotype*
- A string of $\{0,1,2\}^*$ is called a *genotype*
- Two equal length haplotypes generate a unique genotype
- The rules are $0 \oplus 0 = 0$, $1 \oplus 1 = 1$, $0 \oplus 1 = 1 \oplus 0 = 2$
  E.g., $0010, 0101 \Rightarrow 0222$
- If we have a genotype, we can only conjecture (potentially exponentially many) pairs of*haplotype*s that generated it (observe that, e.g., $0110, 0001 \Rightarrow 0222$)
- Biological experiments allow us to know genotypes!
- Since genotypes are introduced in evolution, it is reasonable to find minimal sets of haplotypes explaining the known genotypes.

HAPLOTYPE INFERENCE

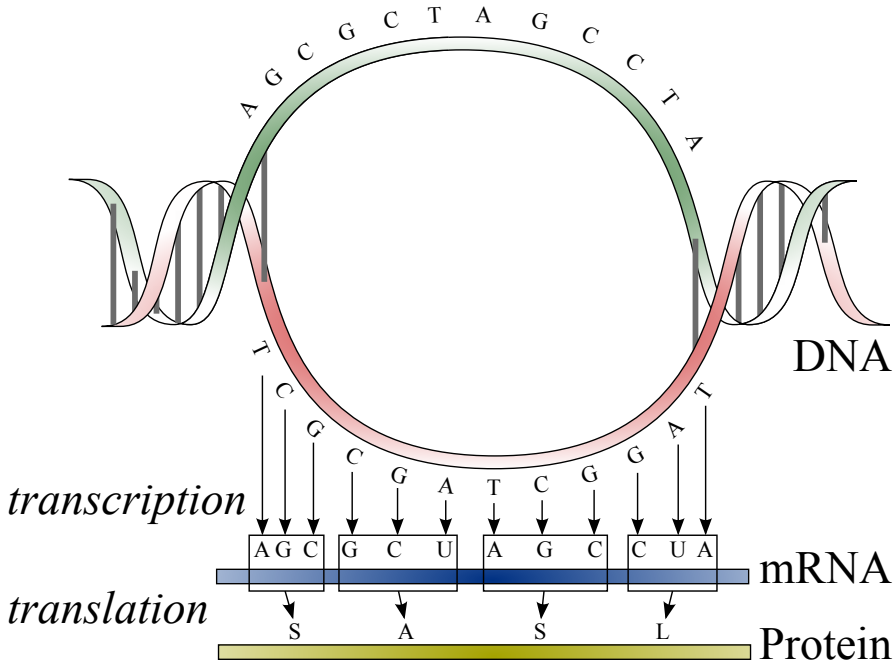
- Let $H \subseteq \{0,1\}^n$ be a set of *haplotypes* and
- $G \subseteq \{0,1,2\}^n$ be a set of *genotypes*.
- Given $h_1, h_2 \in H$ and $g \in G$, $\{h_1, h_2\}$ explains g if and only if $|h_1| = |h_2| = |g|$ and $\forall i \in [1..n]$:

$$\begin{array}{lcl} g[i] \leq 1 & \longrightarrow & h_1[i] = h_2[i] = g[i] \\ g[i] = 2 & \longrightarrow & h_1[i] \neq h_2[i] \end{array}$$

- A set of haplotypes H explains a set of genotypes G if for all $g \in G$ there are $h_1, h_2 \in H$ such that $\{h_1, h_2\}$ explains g.
- Given a set of genotypes G and an integer k, the *haplotype inference problem (HIP)* by pure parsimony is the problem of finding a set H that explains G and such that $|H| = k$ (decision version: NP complete).

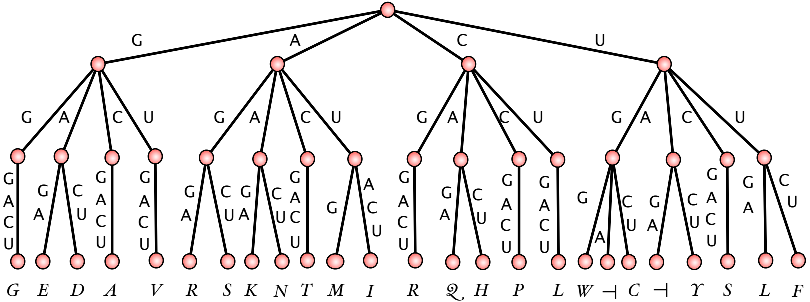
Protein Structure Prediction

PROTEINS AND CENTRAL DOGMA



- The translation phase starts from a mRNA sequence and associates a protein sequence
- Proteins are made of amino acids (20 common different types)
- Amino acids are defined by letters $\{A, \ldots, Z\} \setminus \{B, J, O, U, X, Z\}$

UNIVERSAL CODE

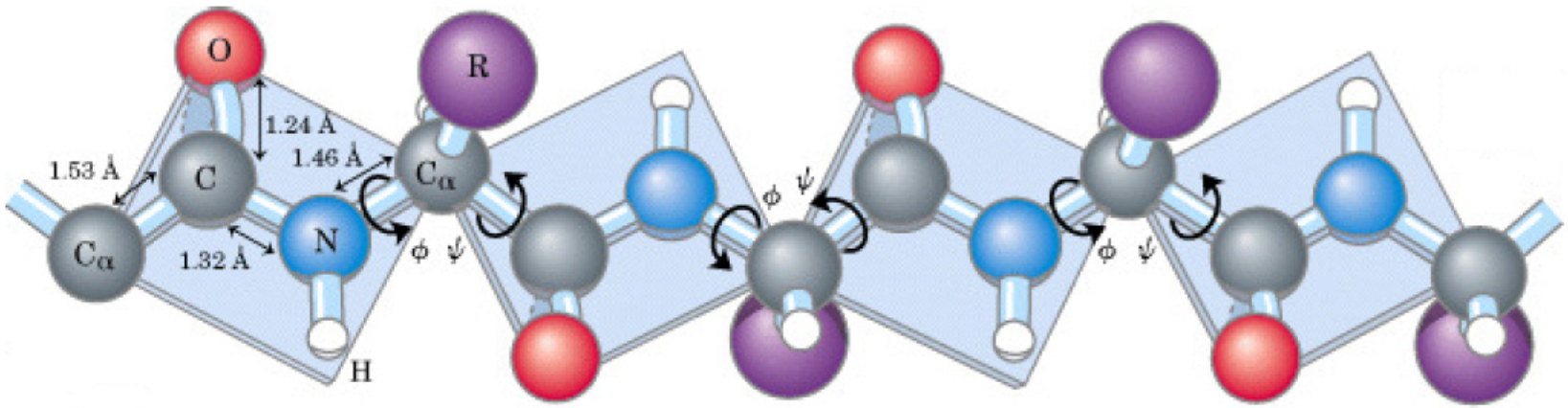


- The translation selects 3 RNA basis and associates 1 amino acid.
- The translation rules are encoded in the universal code.
- The code contains *stop* symbol and some redundant RNA triplets.

Proteins

Amino acids

- Proteins are molecules made of a linear sequence of amino acids.
- Amino acids are combined through *peptide bond*.

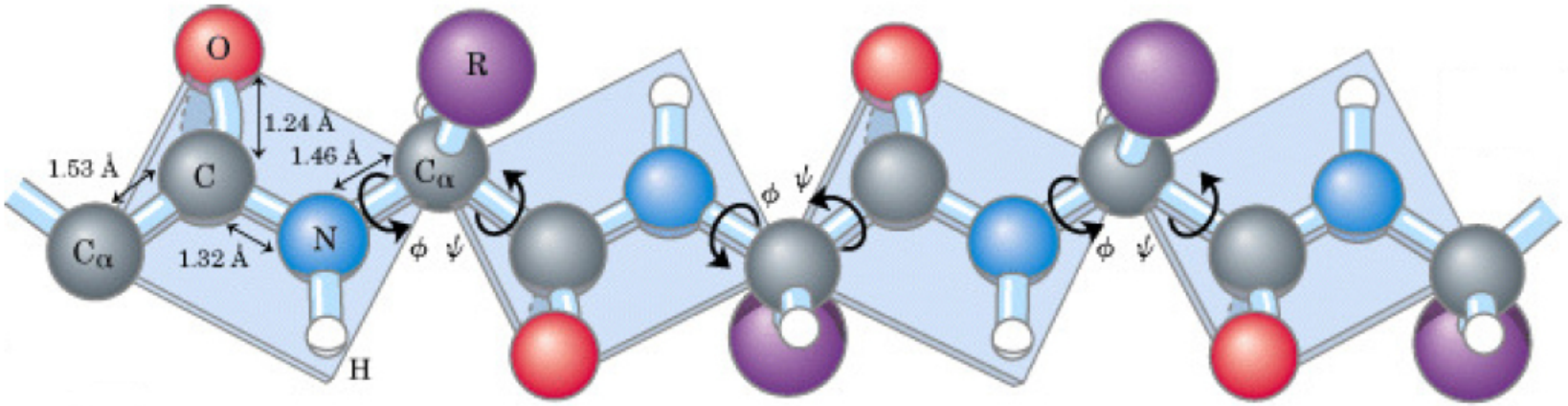


- The purple dots represent the *side chains*, that depend on the amino acid type
- Side chains have different shape, size, charge, polarity, etc.
- A side chain contains from 1 (Glycine) up to 18 (Tryptophan) atoms.

Proteins

Amino acids

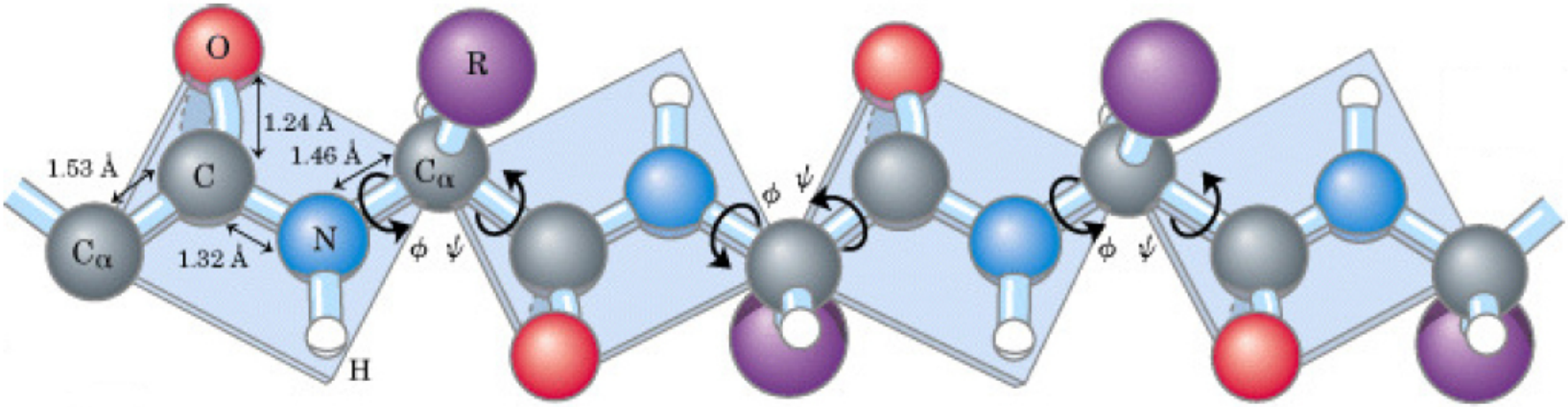
- Proteins are molecules made of a linear sequence of amino acids.
- Amino acids are combined through *peptide bond*.



- The purple dots represent the *side chains*, that depend on the amino acid type
- Side chains have different shape, size, charge, polarity, etc.
- A side chain contains from 1 (Glycine) up to 18 (Tryptophan) atoms.

PROTEINS

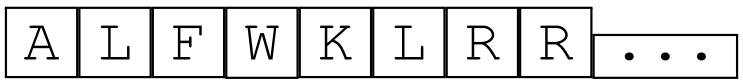
AMINO ACIDS



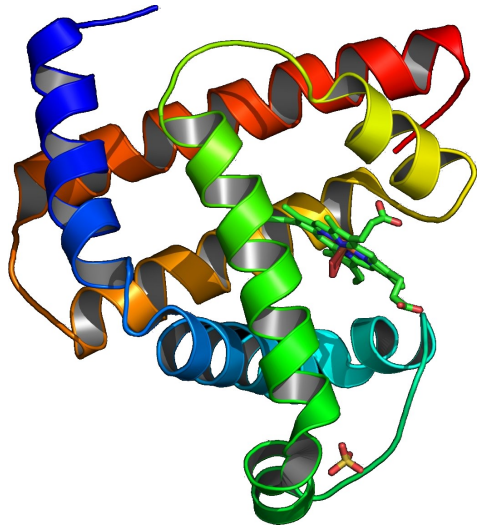
- There are 2 degrees of freedom (black arrows) for each amino acid
- A protein with n amino acids has $2n$ degrees of freedom!
- Typical size range from 50 to 500 amino acids

THE STRUCTURE PREDICTION PROBLEM

- Given the primary structure of a protein (its amino acid sequence)
- For each amino acid, output its position in the space (tertiary structure of a protein)



? ⇓ ?



Proteins

Facts

- Proteins FOLD spontaneously (when in their natural environment)
- Folding is consistent $\Rightarrow$ same protein folds in the same way [Anfinsen74]
- Folding is fast $\Rightarrow$ 1mS – 1S
- Driven by non covalent forces: electrostatic interactions, volume constraints, Hydrogen Bonding, van der Waals, Salt/disulfide Bridges
- Backbone is rigid, interaction with water, ions and ligands

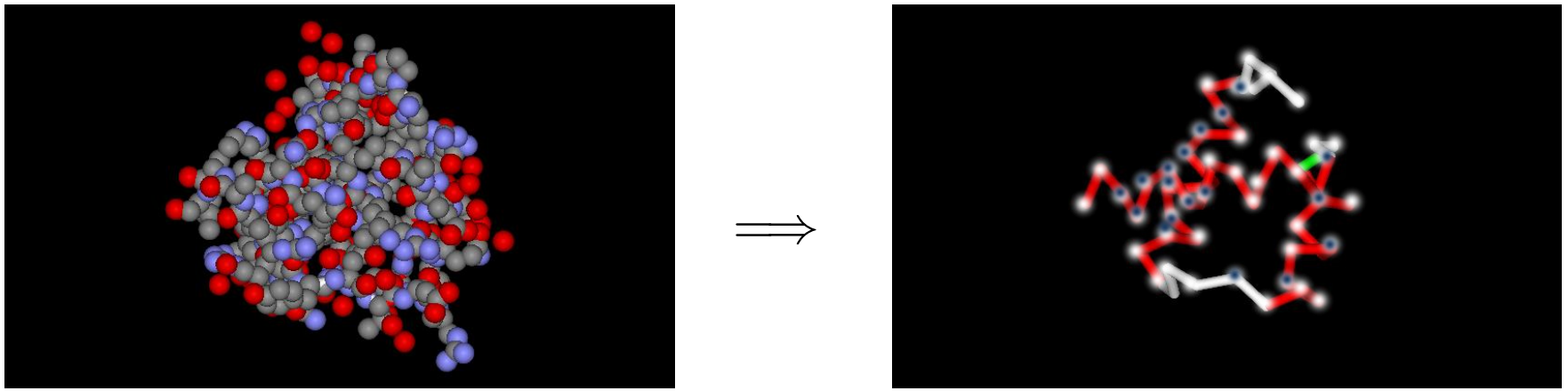
THE STRUCTURE PREDICTION PROBLEM

... and this is the hard part:

- In nature a protein has a unique/stable 3D conformation
- A cost function (that mimics physics laws) can be used to score each conformation
- Searching for the optimal score produces the best candidate is difficult (NP-complete even in extremely simplified modelings)

THE PROTEIN STRUCTURE PREDICTION PROBLEM

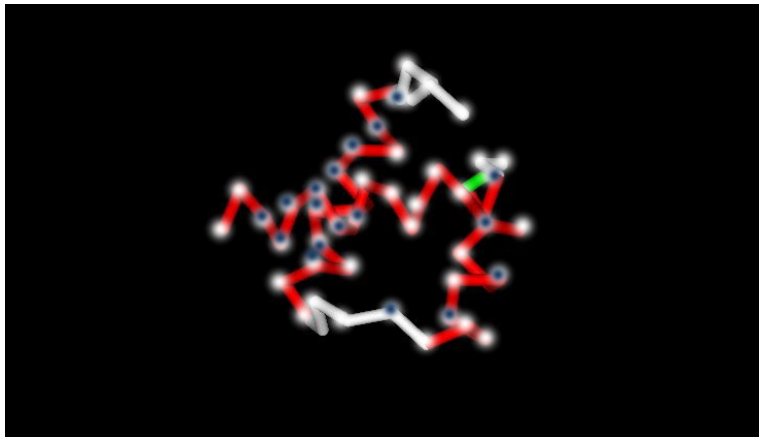
- In this presentation we have two simplifications:
- Protein model: only one atom per amino acid, only 2 classes of amino acids (hydrophobic and polar)



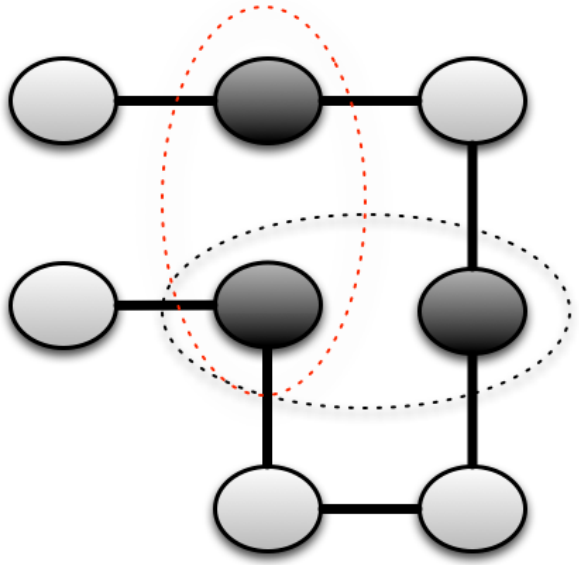
- Spatial model: 2D square lattice to represent amino acid positions

THE PROTEIN STRUCTURE PREDICTION PROBLEM

- In this presentation we have two simplifications:
- Protein model: only one atom per amino acid, only 2 classes of amino acids (hydrophobic and polar)
- Spatial model: 2D square lattice to represent amino acid positions



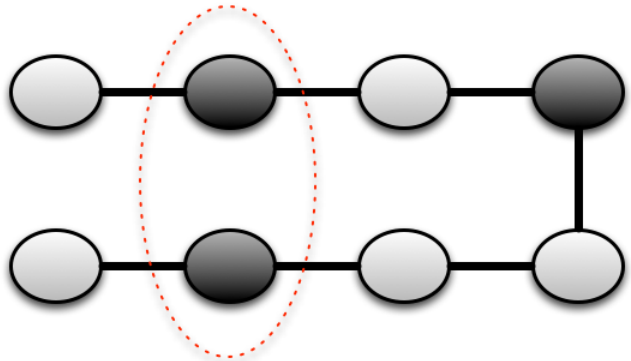
$\Longrightarrow$

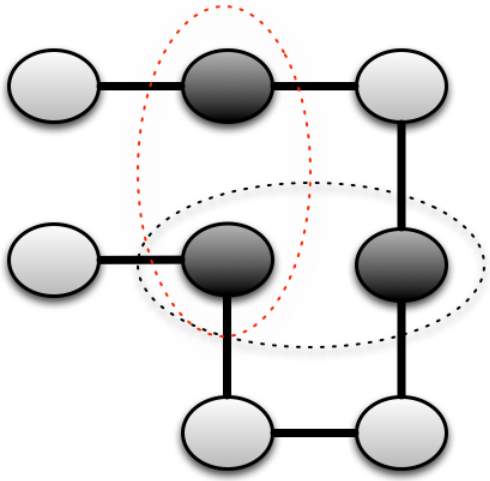


THE PROTEIN STRUCTURE PREDICTION PROBLEM

MODEL

- The input is a list S of amino acids $S = s_1, \ldots, s_n$,
- where $s_i \in \{h, p\}$
- Each s_i is placed on a 2D grid with integer coordinates
- Any pair of two amino acids can't occupy the same position
- If two amino acids are at distance 1, they are in contact





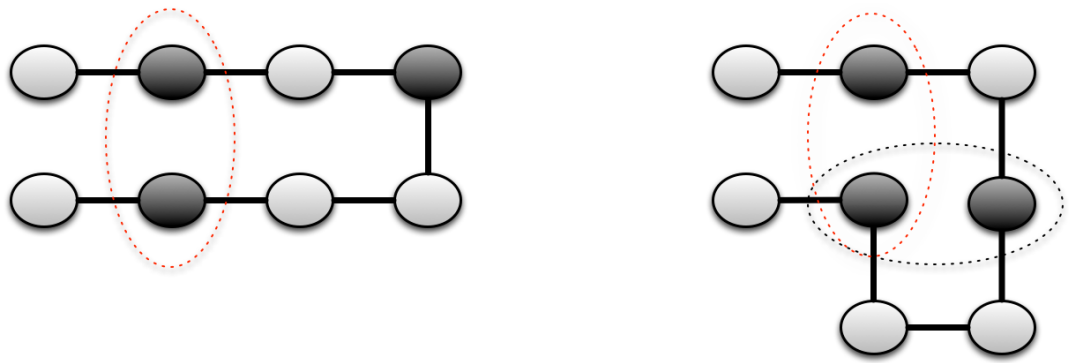
The protein structure prediction problem

Model

- A folding is a function $\omega : \{1, \ldots, n\} \longrightarrow \mathbb{N}^2$ where
- $\forall i \, \text{next}(\omega(i), \omega(i+1))$ and
- $\forall i, j \, (i \neq j \rightarrow \omega(i) \neq \omega(j))$
- $\text{next}(\langle X_1, Y_1 \rangle, \langle X_2, Y_2 \rangle) \Longleftrightarrow |X_1 - X_2| + |Y_1 - Y_2| = 1$.
- Find a folding that minimizes the (simplified) energy function:

$$E(S, \omega) = \sum_{\substack{1 \leq i \leq n-2 \\ i+2 \leq j \leq n}} \text{Pot}(s_i, s_j) \cdot \text{next}(\omega(i), \omega(j))$$

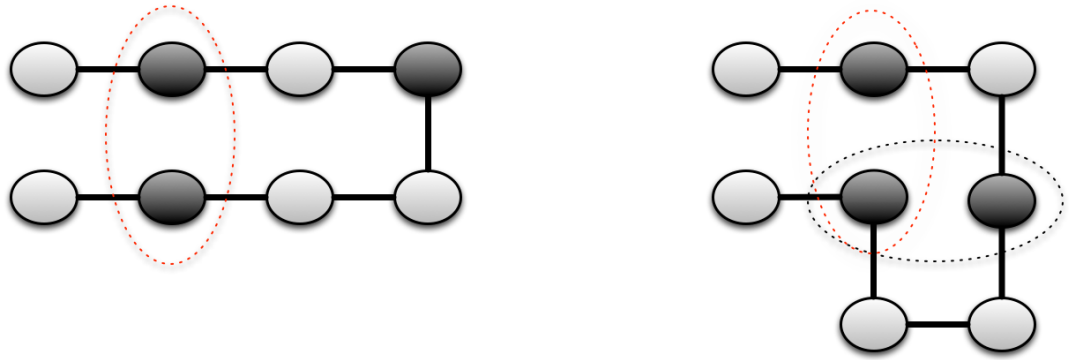
where $\text{Pot}(p, p) = \text{Pot}(h, p) = \text{Pot}(p, h) = 0$ and $\text{Pot}(h, h) = -1$.



- A folding is a function $\omega : \{1, \ldots, n\} \longrightarrow \mathbb{N}^2$ where
- $\forall i \, \text{next}(\omega(i), \omega(i+1))$ and
- $\forall i, j \, (i \neq j \rightarrow \omega(i) \neq \omega(j))$
- $\text{next}(\langle X_1, Y_1 \rangle, \langle X_2, Y_2 \rangle) \Longleftrightarrow |X_1 - X_2| + |Y_1 - Y_2| = 1$.
- Find a folding that minimizes the (simplified) energy function:

$$E(S, \omega) = \sum_{\substack{1 \leq i \leq n-2 \\ i+2 \leq j \leq n}} \text{Pot}(s_i, s_j) \cdot \text{next}(\omega(i), \omega(j))$$

where $\text{Pot}(p, p) = \text{Pot}(h, p) = \text{Pot}(p, h) = 0$ and $\text{Pot}(h, h) = -1$.



- With $\mathbb{N}^2$ and HP, establishing whether there is a folding with energy $< k$ is NP-complete
- (Crescenzi, Goldman, Papadimitriou, Piccolboni, Yannakakis. On the Complexity of Protein Folding. Journal of Computational Biology 5(3): 423-466 (1998))