Exploring Life through Logic Programming

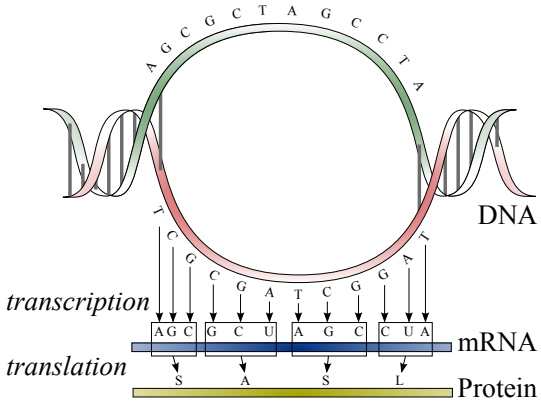
Alessandro Dal Palú Agostino Dovier
Andrea Formisano Enrico Pontelli

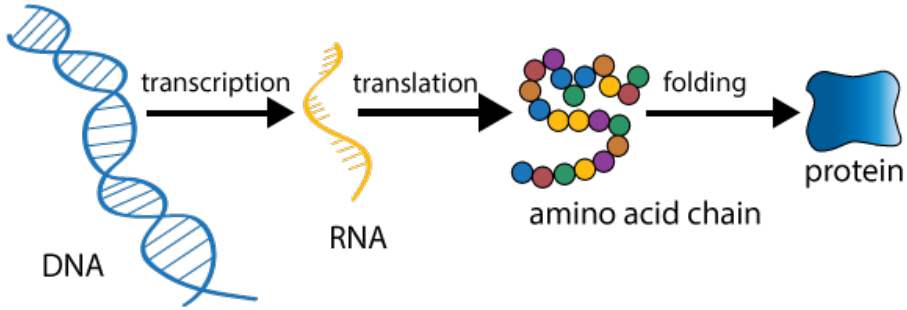
Dept. Computer Science, New Mexico State University, USA

University of Udine, Italy
December 2015

Protein Structure Prediction

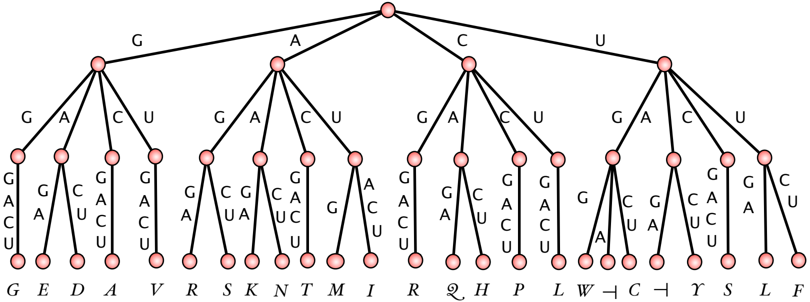
Proteins and Central Dogma





- 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

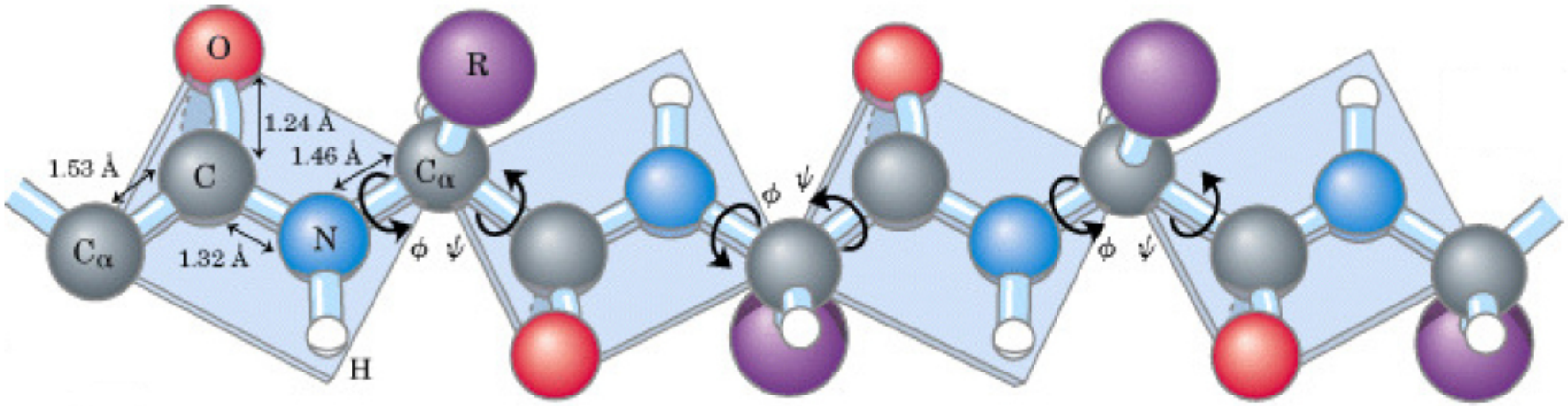


- The translation selects 3 RNA basis (codon) and associates 1 amino acid.
- The translation rules are encoded in the universal code.
- The code contains *stop* symbol and some redundant RNA triplets.

Proteins

Amino acids

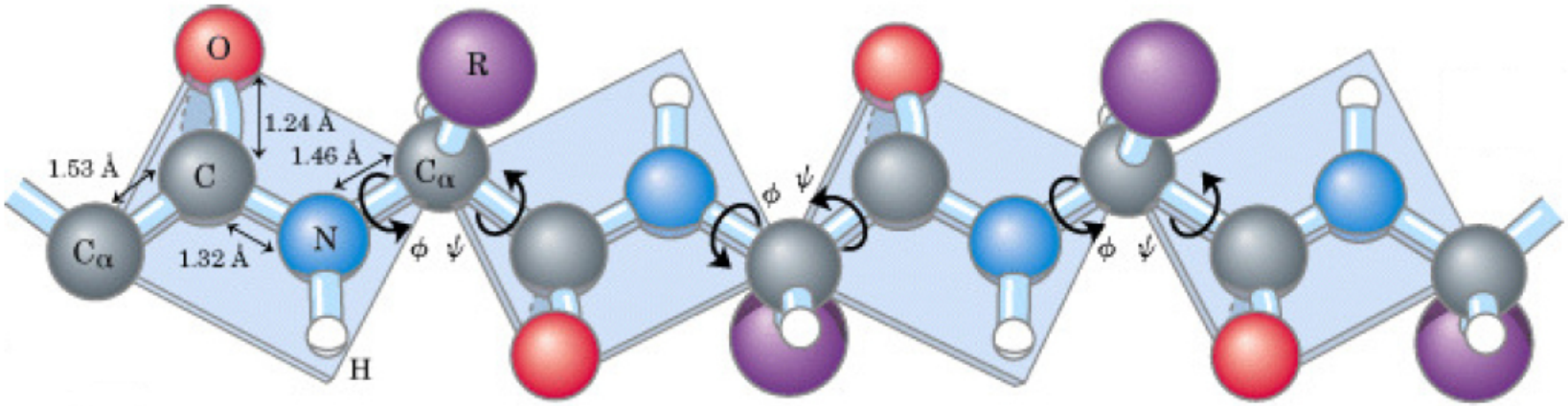
- Proteins are molecules made of a linear sequence of amino acids.
- Amino acids are combined through *peptide bond*.



Proteins

Amino acids

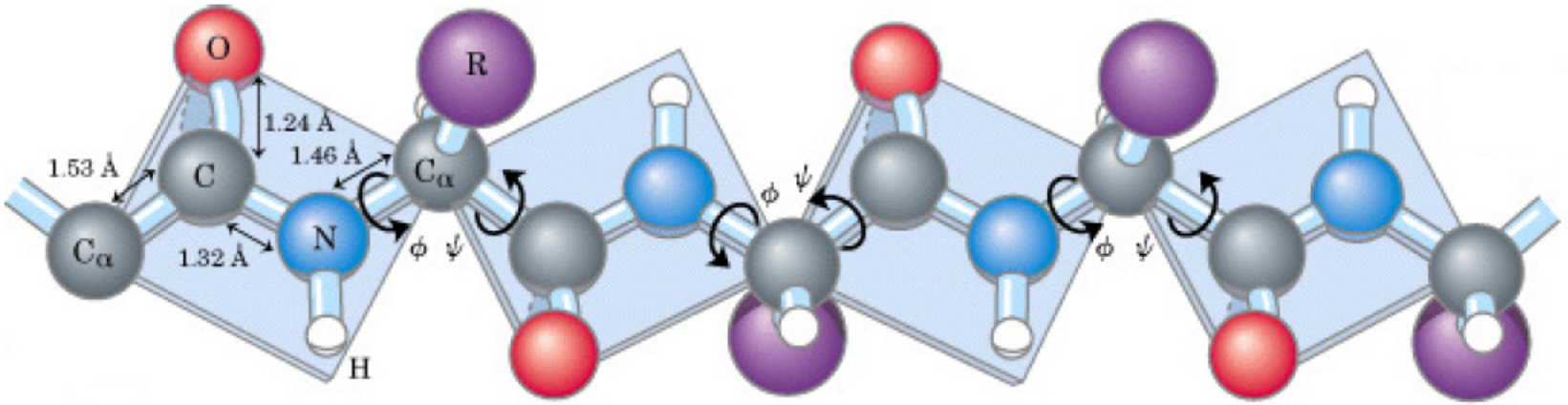
- Proteins are molecules made of a linear sequence of amino acids.
- Amino acids are combined through *peptide bond*.



- 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

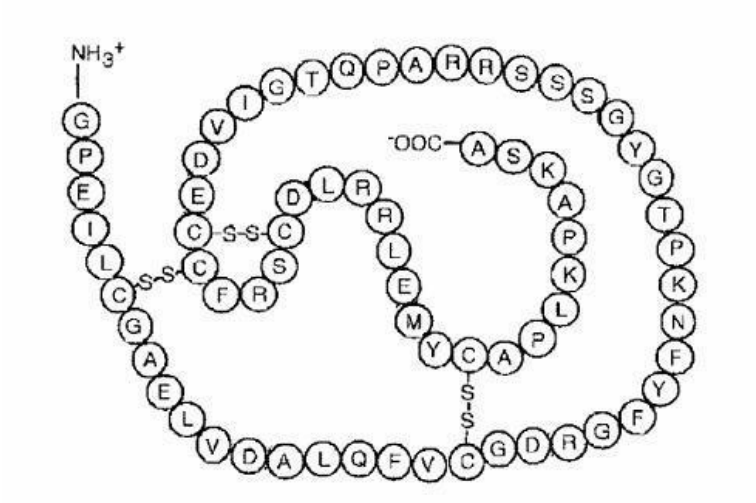


- There are 2 degrees of freedom (black arrows) for each amino acid
- A protein with n amino acids has $2n$ degrees of freedom (ignoring the side chains)
- Typical size range from 50 to 500 amino acids

Proteins

Primary sequence

Amino acid sequence:
GPEILCGAELVDALQFVCGDRGFYFNKPTGYGSSS
RRAPQTGIVDECCFRSCDLRRLEMYCAPLKPAKSA

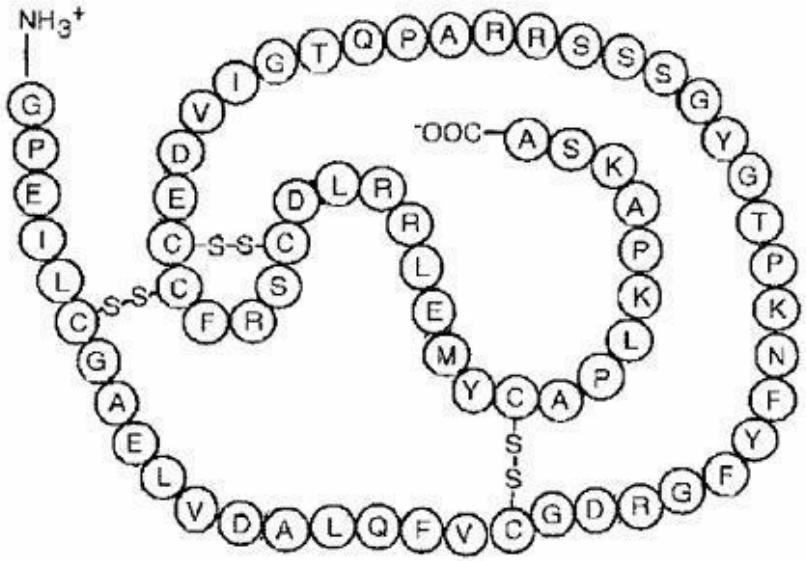


Primary sequence...

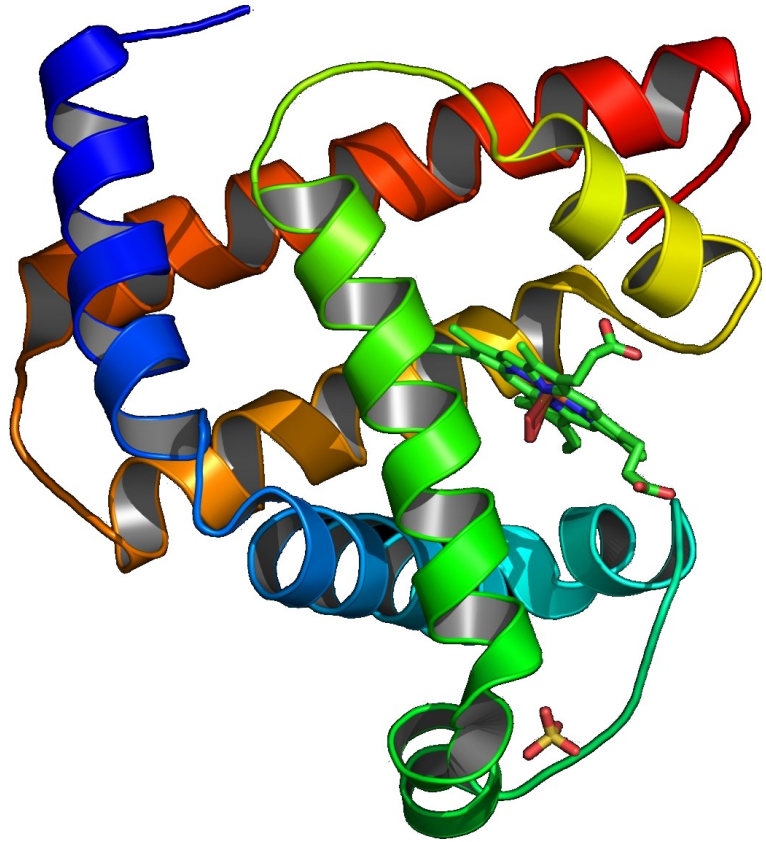
...embedded in the 3D space

Proteins

Tertiary structure



Primary sequence...



... tertiary structure

- Proteins FOLD spontaneously (when in their natural environment)

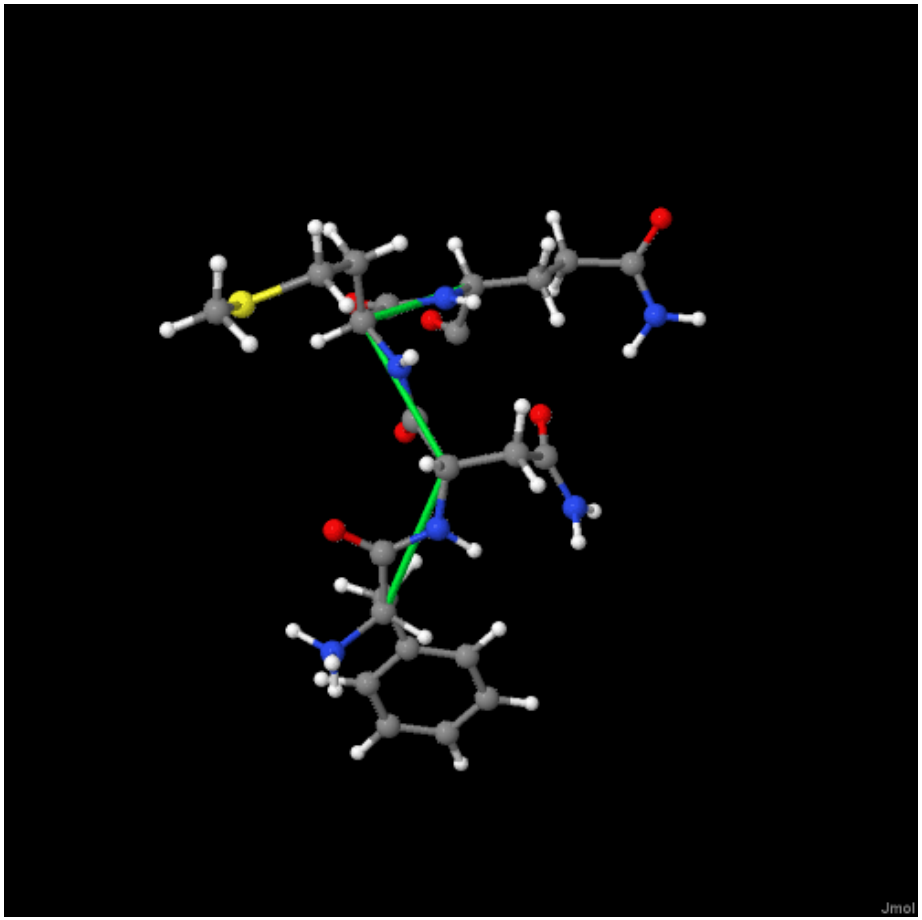
Proteins

Facts

- Folding is consistent $\Rightarrow$ same protein folds in the same way [Anfinsen74]
- Folding is fast $\Rightarrow$ 1μS – 1mS
- Driven by non covalent forces: electrostatic interactions, volume constraints, Hydrogen Bonding, van der Waals, Salt/disulfide Bridges
- Backbone is rigid

Proteins

Simplified structure

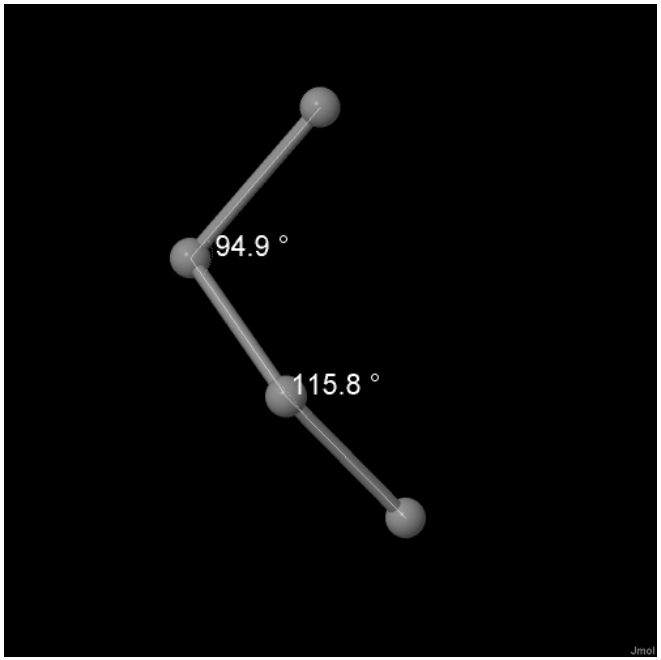


- The *backbone* (green) links consecutive amino acids
- Each amino acid can be identified by its carbon alpha (C_α)
- The distance between consecutive $C\alpha$ points is constant (3.81Å).

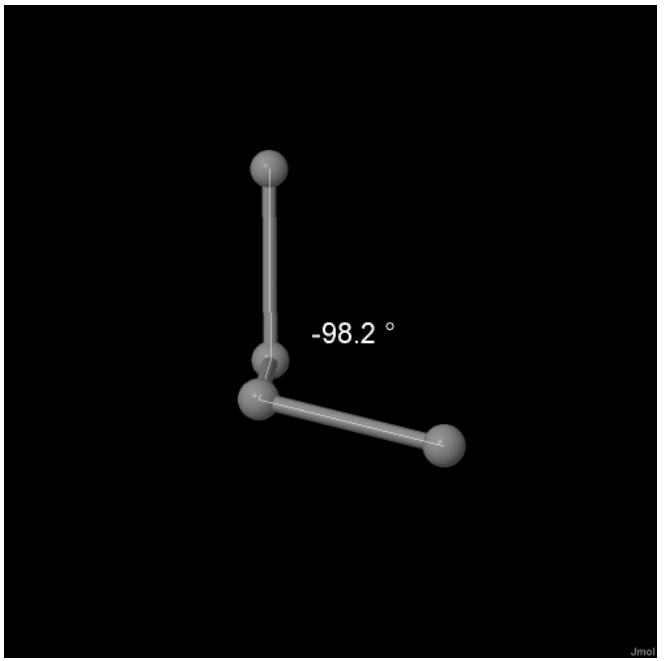
Proteins

Backbone

- The distance between $C\alpha$ points is constant (3.81Å).
- Let us consider a backbone made of 4 amino acids.



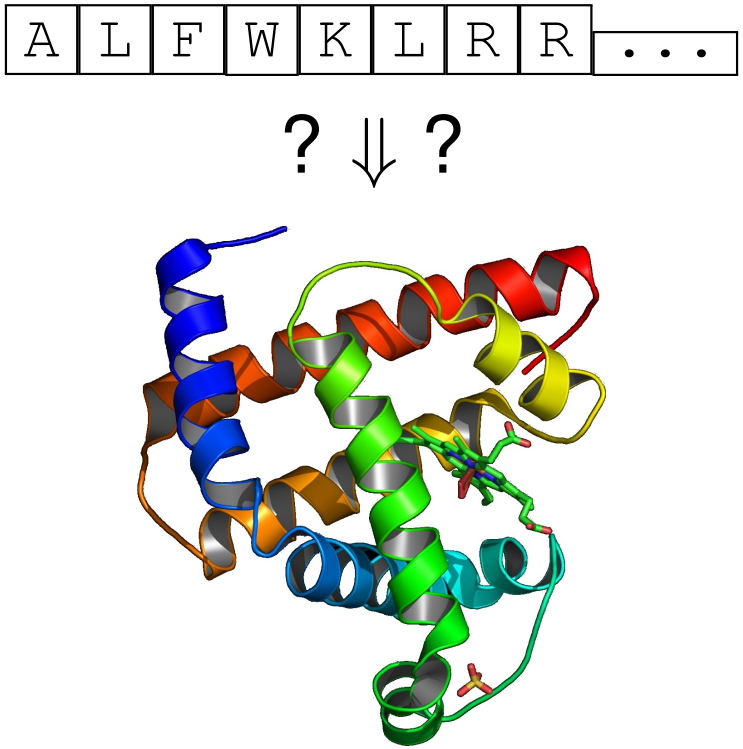
There are two bend angles



And one torsional angle.

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)



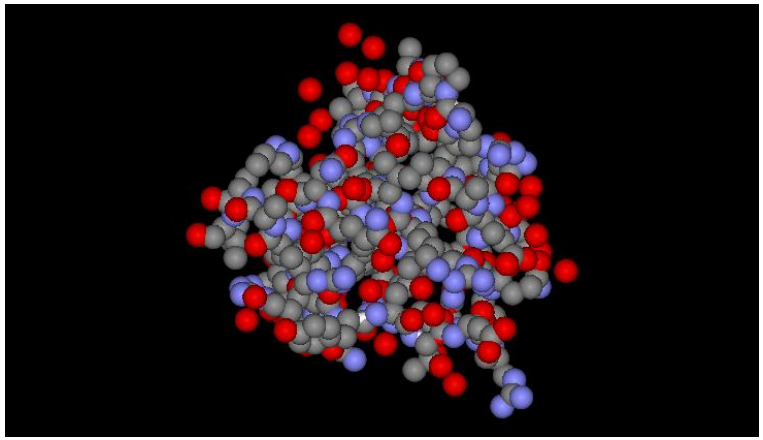
The structure prediction problem

... and this is the hard part:

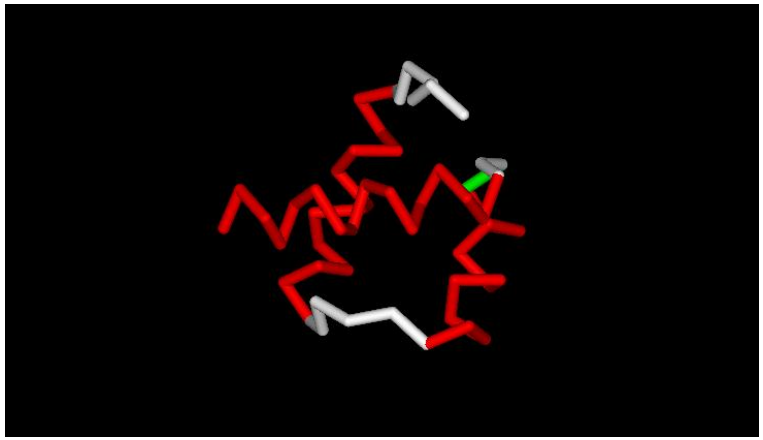
- In nature a protein has a unique/stable 3D conformation
- A cost function (that mimics physics laws) can be used to score each conformation
- Searching for the optimal score produces the best candidate is difficult (NP-complete even in extremely simplified modelings)

The protein structure prediction problem

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

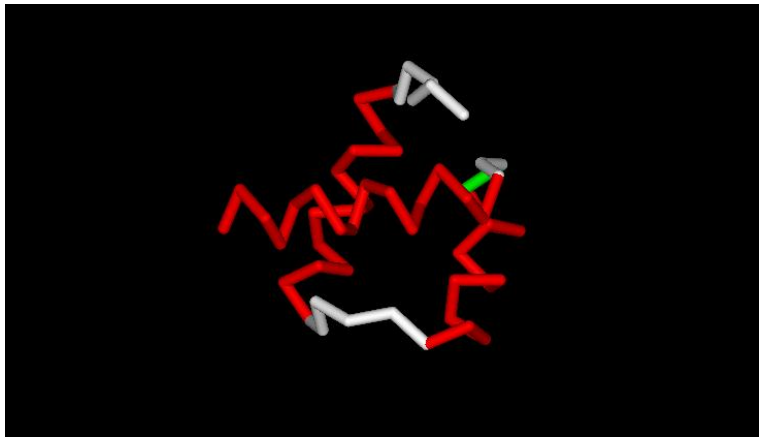


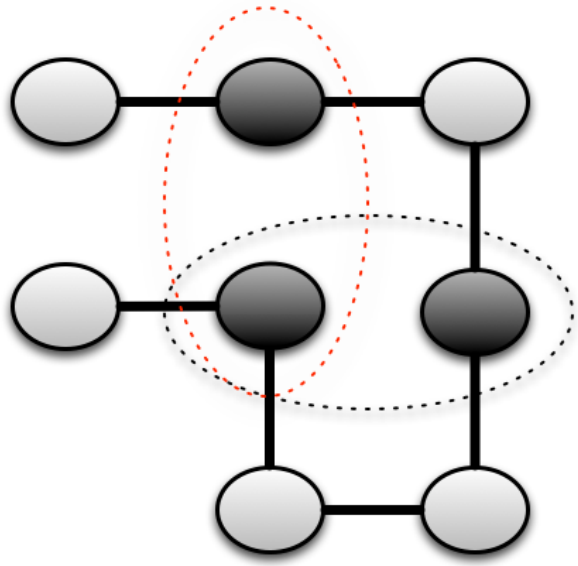
$\Longrightarrow$



The protein structure prediction problem

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

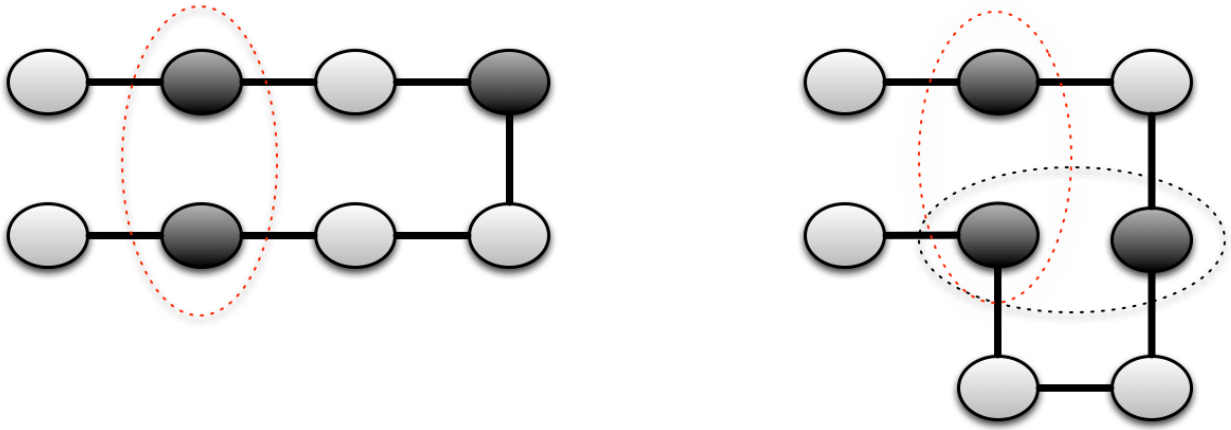




The protein structure prediction problem

Model

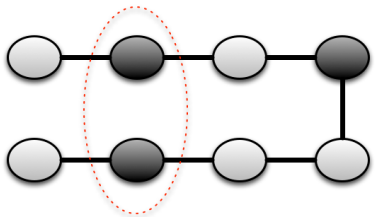
- The input is a list S of amino acids $S = s_1, \ldots, s_n$, where $s_i \in \{h, p\}$
- Each s_i is placed on a 2D grid with integer coordinates
- Any pair of two amino acids cannot occupy the same position
- If two amino acids are at distance 1, they are in contact

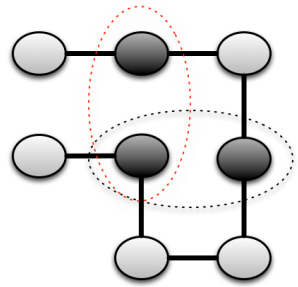


The protein structure prediction problem

Model

- Let: $\text{next}(\langle X_1, Y_1\rangle, \langle X_2, Y_2\rangle) \iff |X_1 - X_2| + |Y_1 - Y_2| = 1$.
- A folding is a function $\omega : \{1, \ldots, n\} \longrightarrow \mathbb{N}^2$ where:
  - For each $1 \leq i < n$: $\text{next}(\omega(i), \omega(i+1))$
  - For each $1 \leq i, j \leq n$: $(i \neq j \rightarrow \omega(i) \neq \omega(j))$





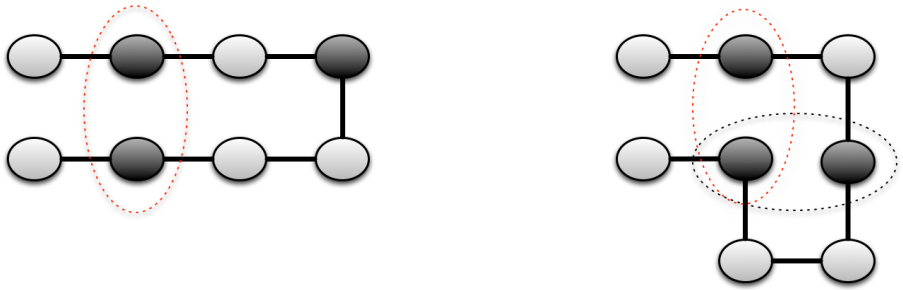
The protein structure prediction problem

Model

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

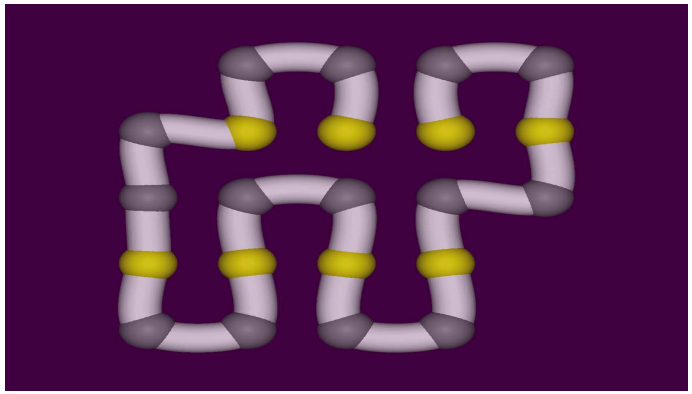
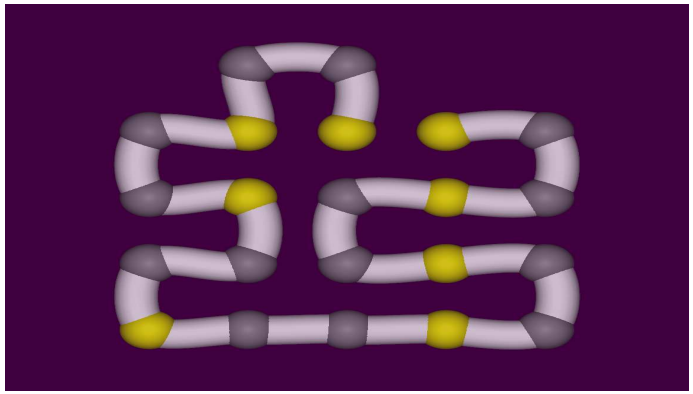
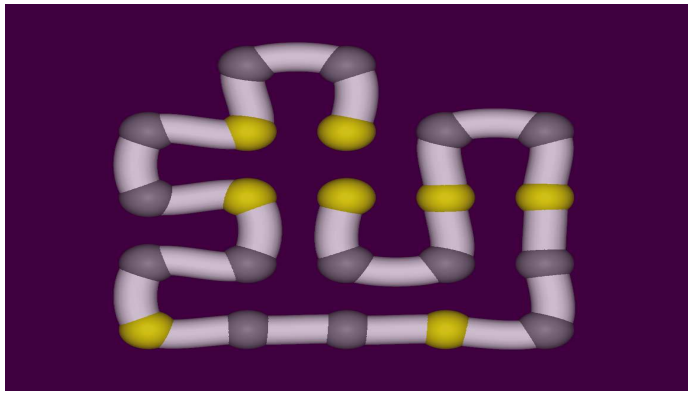
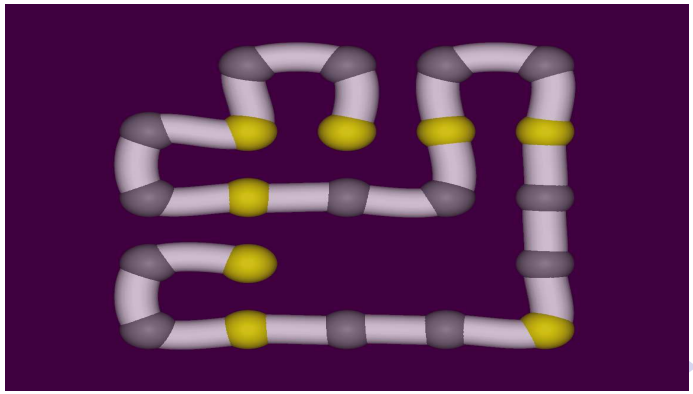
$$E(S, \omega) = \sum_{\substack{1 \leq i \leq n-2 \\ i+2 \leq j \leq n}} \text{Pot}(s_i, s_j) \cdot \text{next}(\omega(i), \omega(j))$$

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



Model

- H: Yellow; P: Grey
- All foldings have energy -6

The protein structure prediction problem

Complexity

- Establishing whether there is a folding with energy $< k$ in the HP model and $\mathbb{N}^2$ space is NP-complete

  (Crescenzi, Goldman, Papadimitriou, Piccolboni, Yannakakis. On the Complexity of Protein Folding. Journal of Computational Biology 5(3): 423-466 (1998))

The protein structure prediction problem

ASP Model

- The positions can be restricted to $[1 \ldots 2n] \times [1 \ldots 2n]$
- The first two amino acids are set to $\omega(1) = \langle n, n \rangle$ and $\omega(2) = \langle n, n+1 \rangle$
- Sequence is represented by n facts:

```
prot(1,p).  prot(2,h).  prot(3,p).
prot(4,h).  prot(5,p).  prot(6,p).
prot(7,h).  prot(8,p).
```

- Each placement is stored as `sol(i,X,Y).`

ASP Encoding

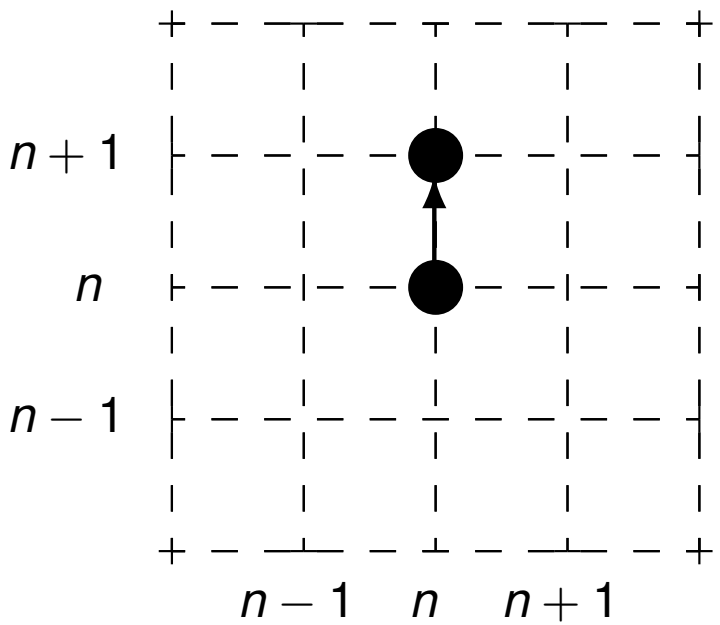
```
(1)   size(N)          :- N = #count{ I:prot(I,J) }.
(2)   board(1..2*N)    :- size(N).
(3)   range(X,Y)       :- size(N), board(X), board(Y),
                          |X-N|+|Y-N| < N.
```

- Count the protein length and defines the board size
- `range` is loaded with all combinations from 1 to 2N

ASP Encoding

```
(1)  size(N)          :- N = #count{ I:prot(I,J) }.
(2)  board(1..2*N)    :- size(N).
(3)  range(X,Y)       :- size(N), board(X), board(Y),
                         |X-N|+|Y-N| < N.

(4)  sol(1,N,N)       :- size(N).
(5)  sol(2,N,N+1)     :- size(N).
```

- Set the first two amino acids positions



ASP Encoding

```
(1)  size(N)          :- N = #count{ I:prot(I,J) }.
(2)  board(1..2*N)    :- size(N).
(3)  range(X,Y)       :- size(N), board(X), board(Y),
                         |X-N|+|Y-N| < N.

(4)   sol(1,N,N)      :- size(N).
(5)   sol(2,N,N+1)    :- size(N).

(6)   1 { sol(I,X,Y) : range(X,Y) } 1 :- prot(I,Amino).
```

- Each amino acid *I* is associated to a unique position

ASP Encoding

```
(1)  size(N)          :- N = #count{ I:prot(I,J) }.
(2)  board(1..2*N)    :- size(N).
(3)  range(X,Y)       :- size(N), board(X), board(Y),
                         |X-N|+|Y-N| < N.

(4)   sol(1,N,N)      :- size(N).
(5)   sol(2,N,N+1)    :- size(N).

(6)   1 { sol(I,X,Y) : range(X,Y) } 1 :- prot(I,Amino).

(7)   :- prot(I1,A1), prot(I2,A2), I1<I2,
         sol(I1,X,Y), sol(I2,X,Y),  range(X,Y).
```

- Constraint: two distinct amino acids cannot overlap

ASP Encoding

```
(8)   :- prot(I1,A1), prot(I2,A2),  I2>1,
                I1==I2-1, not next(I1,I2).
(9)    next(I1,I2) :- prot(I1,A1), prot(I2,A2), I1<I2,
                sol(I1,X1,Y1), sol(I2,X2,Y2),
                range(X1,Y1), range(X2,Y2),
                1 == |X2-X1| + |Y2-Y1|.
```

- Constraint: two consecutive amino acids in sequence are `next` in space

ASP Encoding

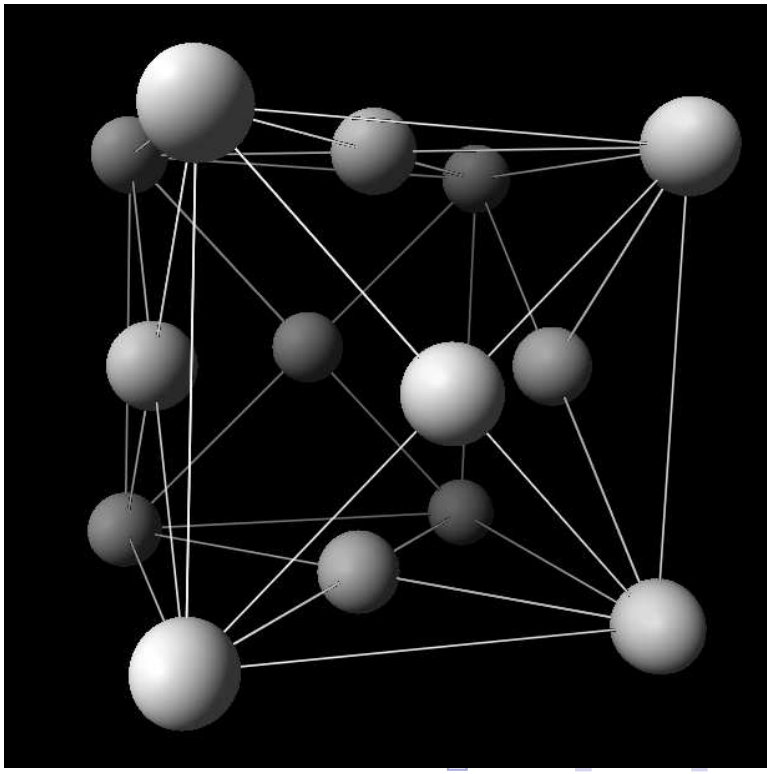
```
(8)   :- prot(I1,A1), prot(I2,A2),  I2>1,
                I1==I2-1, not next(I1,I2).
(9)    next(I1,I2) :- prot(I1,A1), prot(I2,A2), I1<I2,
                sol(I1,X1,Y1), sol(I2,X2,Y2),
                range(X1,Y1), range(X2,Y2),
                1 == |X2-X1| + |Y2-Y1|.

(10)  energy_pair(I1,I2) :- prot(I1,h), prot(I2,h),
                 next(I1,I2), |I1-I2|>2,
                 1==(I2-I1)  \ 2. %%% "\" is "mod"
(11)  energy(N) :- N = #count{(I1,I2):energy_pair(I1,I2)}.
(12)  #maximize { N:energy(N)}.
```

- If two amino acids of type *h* are `next`, count one contact (energy_pair)
- Optimization: at least one amino acid in sequence between the pair
- Optimization: only odd distances in space can contribute to a contact

Extensions

- Generalization to 3d cubic space is trivial.
- organization of the space have been explored
- The Face Centered Cube lattice models the discrete space in which the protein can fold.

- Two points are *connected* (next) iff $|x_i - x_j|^2 + |y_i - y_j|^2 + |z_i - z_j|^2 = 2$,
- Each point has 12 neighbors



Some References

- R. Backofen and S. Will, A constraint-based approach to fast and exact structure prediction in 3-dimensional protein models, Constraints 11(1):5-30, 2006.
- A. Dal Palù, A. Dovier and F. Fogolari. Constraint logic programming approach to protein structure prediction, BMC Bioinformatics 5(186), 2004.
- A. Dal Palù, A. Dovier and E. Pontelli, A constraint solver for discrete lattices, its parallelization, and application to protein structure prediction, Software Practice and Experience 37(13):1405-1449, 2007. (COLA)
- A. Dal Palù, A. Dovier and E. Pontelli. Computing approximate solutions of the protein structure determination problem using global constraints on discrete crystal lattices, Int'l Journal of Data Mining and Bioinformatics 4(1):1–20, 2010. Also in WCB 06 and WCB 07
- P. Barahona and L. Krippahl, Constraint programming in structural bioinformatics, Constraints 13(1-2):3-20, 2008.
- A. Dovier. Recent constraint/logic programming based advances in the solution of the protein folding problem. Intelligenza Artificiale 5(1):113-117, 2011.
- Approximated results with local search and/or LNS by Hoos et al. and by Van Hentenryck et al.