Constraint Programming e sue applicazioni alla Predizione di Struttura

Agostino Dovier

Università di Udine
Dipartimento di Matematica e Informatica

Udine, 16 Maggio 2008

CP e PSP

Agostino Dovier

Richiami sul Protein Folding

Codifica

Evoluzione

Sommario

Richiami sul Protein Folding

Codifica

Evoluzione

CP e PSP

Agostino Dovier

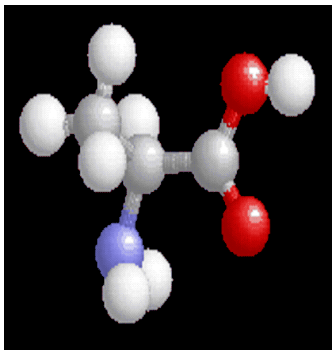
Richiami sul Protein Folding

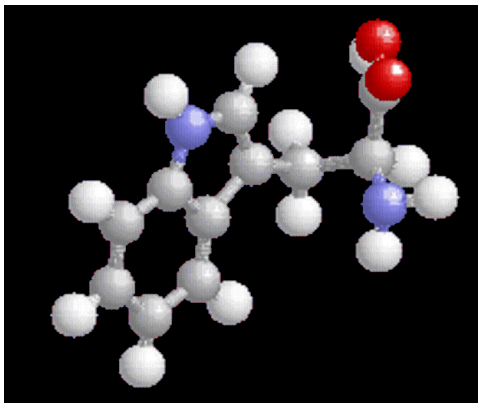
Codifica

Evoluzione

- La struttura primaria di una proteina è una sequenza di monomeri, detti aminoacidi, di lunghezza variabile da poche unità ad alcune centinaia.
- Gli aminoacidi sono di 20 tipi: **Ala**nine (A), **Cys**teine (C), **Asp**artic Acid (D), **Glu**tamic Acid (E), **Phe**nylalanine (F), **Gly**cine (G), **His**tidine (H), **Iso**l**e**ucine (I), **Lys**ine (K), **Leu**cine (L), **Met**hionine (M), **As**paragi**n**e (N), **Pro**line (P), **Gl**utami**n**e (Q), **Arg**inine (R), **Ser**ine (S), **Thr**eonine (T), **Val**ine (V), **Tryp**tophan (W), **Tyr**osine (Y).
- Per noi la struttura primaria di una proteina è una stringa di lettere nell'alfabeto $\{A, \ldots, Z\} \setminus \{B, J, O, U, X, Z\}$.

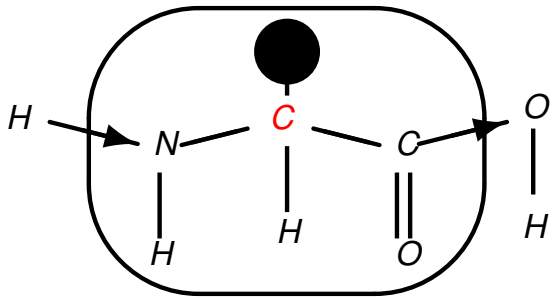
Il problema del protein folding
Due aminoacidi (Alanina, Triptofano)





$C_3H_7NO_2 \rightarrow 13$ atoms $\quad C_{11}H_{12}N_2O_2 \rightarrow 27$ atoms

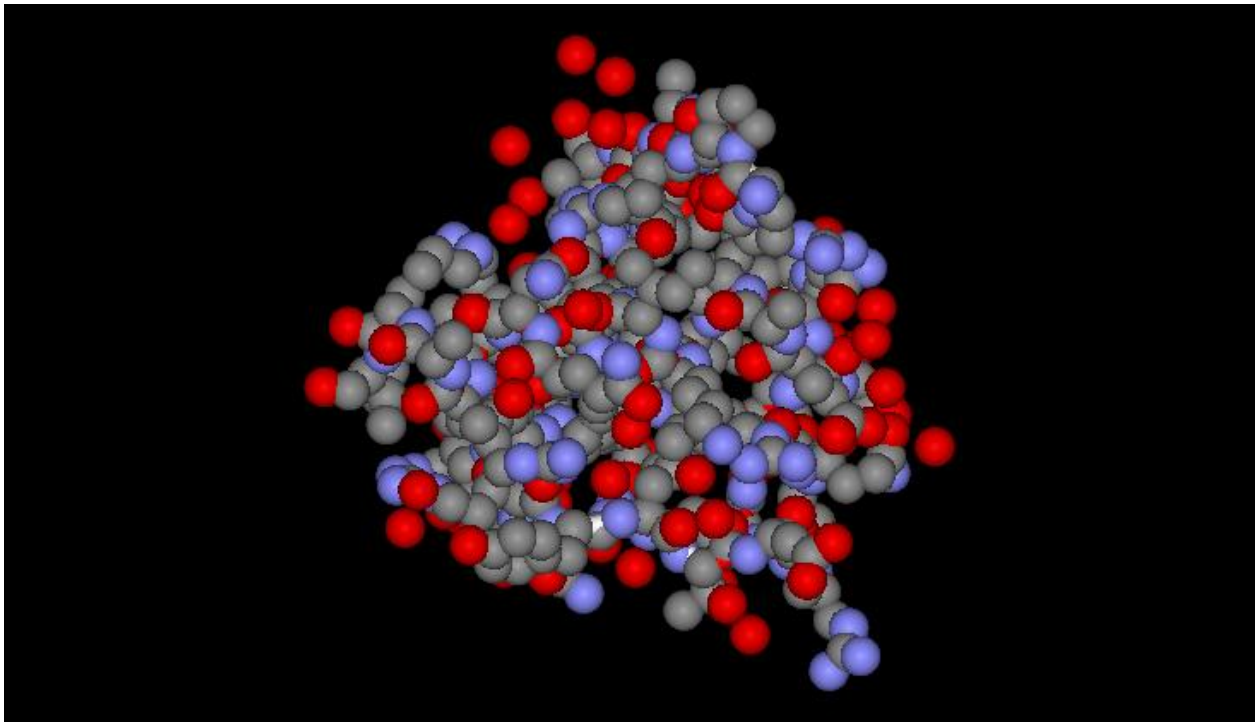
White $=$ H
Blue $=$ N
Red $=$ O
Grey $=$ C



Il problema del protein folding
La struttura terziaria

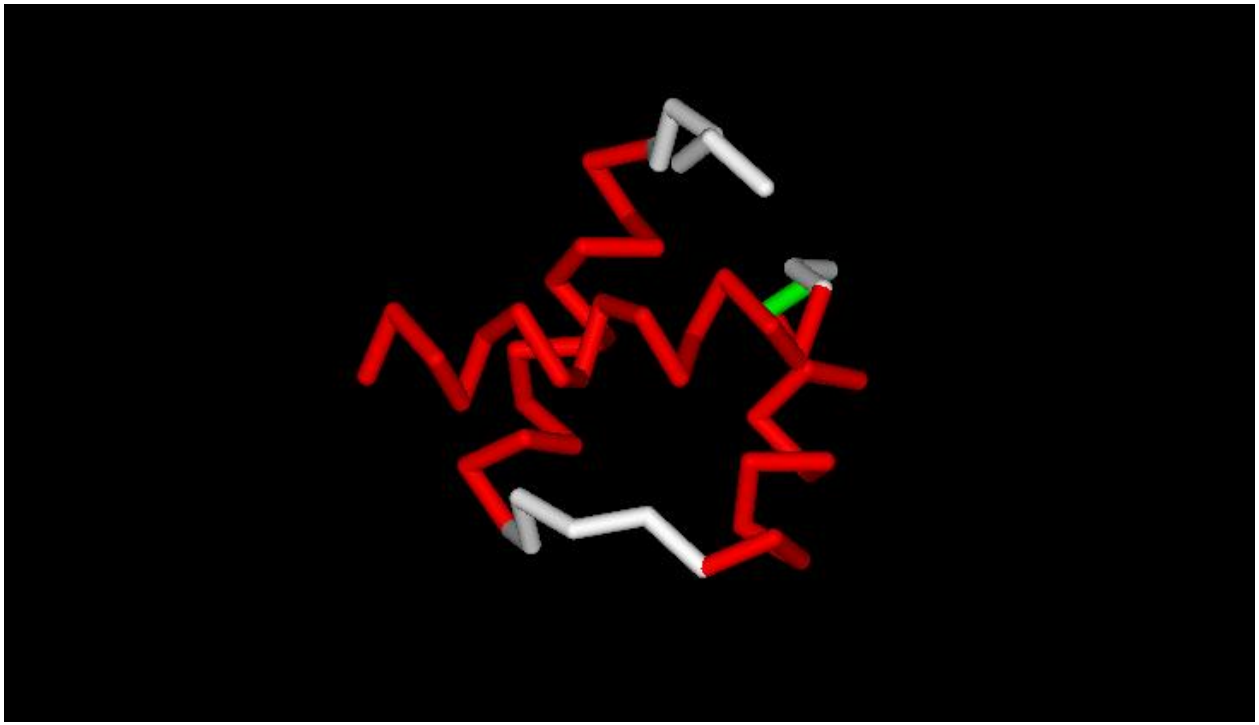
- La struttura terziaria è la forma 3D che assume la proteina.
- La proteina si folda a condizioni di temperatura standard raggiungendo l'unico stato nativo.
- Tale stato è quello che minimizza l'energia libera della proteina (*Anfinsen*'s hypothesis)
- e determina la funzione della proteina.
- Il problema del protein structure prediction è il problema di predire la struttura terziaria data la struttura primaria
- Con protein folding si indica anche il processo necessario per raggiungere la struttura terziaria.

Il problema del protein folding
Struttura terziaria di 1ENH



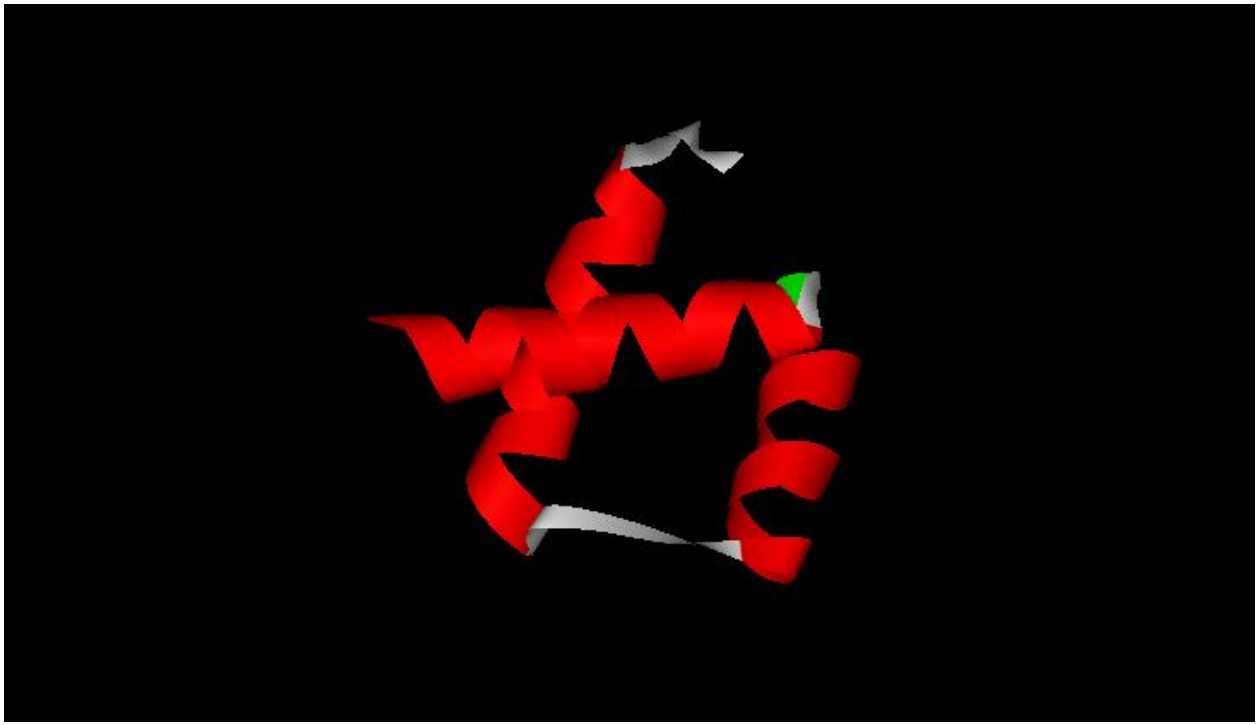
Terziaria Completa

Il problema del protein folding
Struttura terziaria di 1ENH



Sequenza dei $C\alpha$ — distanza costante!

Il problema del protein folding
Struttura terziaria di 1ENH

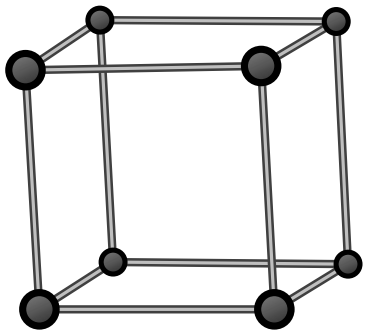
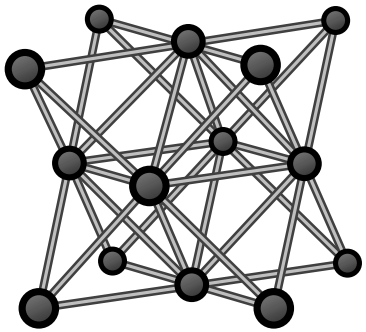
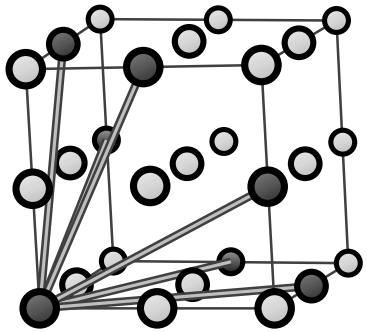


Evidenziate le eliche

The PSP Problem

- Anfinsen: the native state minimizes the whole protein energy. Two problems emerge.
- Energy model:
  - What is the energy function $\mathbb{E}$?
  - It depends on what?
- Spatial Model: Assume $\mathbb{E}$ be known, depending on the aminoacids $a_1, \ldots, a_n$ and on their positions, what is the search's space where looking for the conformation minimizing $\mathbb{E}$?
  - Lattice (discrete) models.
  - Off-lattice (continuous) models.
- Given a choice for spatial and energy models, we can try to study and solve the minimization problem
- If the solution's space is finite, a brute-force algorithm would be, in principle, possible.

Discrete Models
Cubic, FCC, Chess Knight

The PSP as a minimization problem

- Let us define the minimization problem, under the *assumption* that each aminoacid is considered as a whole: a sphere centered in its $C\alpha$-atom.
- The distance between two consecutive $C\alpha$ atoms is fixed (3.8Å).
- Let $\mathcal{L}$ be the set of admissible points for each aminoacid.
- Given the sequence $a_1 \ldots a_n$, a *folding* is a function

$$\omega : \{1, \ldots, n\} \longrightarrow \mathcal{L}$$

such that:

  - next$(\omega(i), \omega(i+1))$ for $i = 1, \ldots, n-1$, and
  - $\omega(i) \neq \omega(j)$ for $i \neq j$.

Objective function

- *Assumption*: the energy is the sum of the energy contributions of each pair of non-consecutive aminoacids.
- It depends on their distance and on their type. The contribution is of the form en_contrib(ω, i, j).
- The function to be minimized is therefore:

$$E(\omega) = \sum_{\substack{1 \leq i \leq n \\ i+2 \leq j \leq n}} \text{en\_contrib}(\omega, i, j)$$

- It is a constrained minimization problem (recall that: next$(\omega(i), \omega(i+1))$ and $\omega(i) \neq \omega(j)$).
- It is parametric on $\mathcal{L}$, next, and en_contrib.
- next and en_contrib are typically non linear.

A first proposal for the Energy: DILL

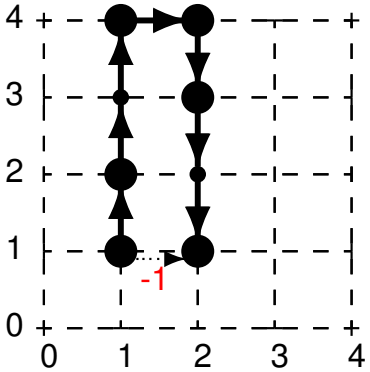
- The aminoacids: Cys (C), Ile (I), Leu (L), Phe (F), Met (M), Val (V), Trp (W), His (H), Tyr (Y), Ala (A) are *hydrophobic* (H).
- The aminoacids: Lys (K), Glu (E), Arg (R), Ser (S), Gln (Q), Asp (D), Asn (N), Thr (T), Pro (P), Gly (G) are *polar* (P).
- The protein is in water: hydrophobic elements tend to occupy the center of the protein.
- Consequently, H aminoacids tend to stay close each other.
- polar elements tend to stay in the frontier.

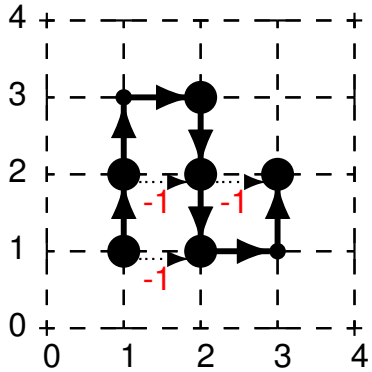
A FIRST PROPOSAL FOR THE ENERGY: DILL

- This fact suggest an energy definition: if two aminoacids of type H are *in contact* (i.e. no more distant than a certain value) in a folding they contribute negatively to the energy.
- The aminoacid is considered as a whole: a unique sphere centered in its $C\alpha$ atom.
- The notion of being *in contact* is naturally formalized in lattice models: one (or more) *lattice units*.

The simplest PFP formalization

- The spatial model is a subset of $\mathbb{N}^2$.
- A *contact* is when $|X_1 - X_2| + |Y_1 - Y2| = 1$.
- The primary list is a sequence of h and p.
- Each contact between pairs of h contributes as -1.
- We would like to find the folding(s) minimizing this energy
- Example: $[h, h, p, h, h, h, p, h]$





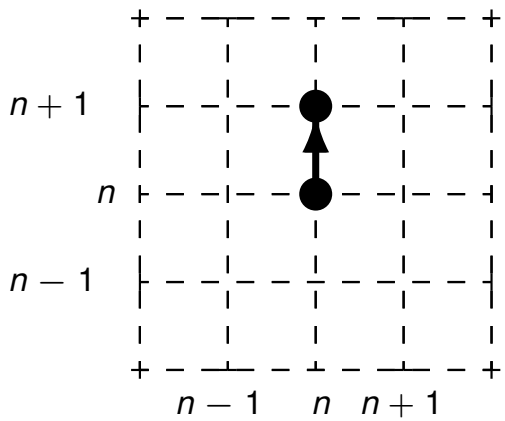
- Unfortunately, the decision version: Is there a folding with Energy $< k$? is NP-complete

HP on $\mathbb{N}^2$

- If the primary structure is $[a_1, \ldots, a_n]$ with $a_i \in h, p$, then

$$\omega(i) \in \mathcal{L} = \{(i,j) : i \in [1..2n-1], j \in [1..2n-1]\}$$

- W.l.o.g, we can assume that $\omega(1) = (n, n)$.
- To avoid simple symmetries, w.l.o.g., we can assume that $\omega(2) = (n, n+1)$.



- We need to implement *next*, *en_contrib*, ...
- Let us see a simple (and working) $CLP(\mathcal{FD})$ code.

Encoding the PF, HP, on $\mathbb{N}^2$


```
%%% Primary = [a1,...,aN], ai in {h,p}
pf(Primary, Tertiary) :-
        constrain(Primary,Tertiary,Energy),
        labeling([ff,minimize(Energy)],Tertiary).
%%% Tertiary = [X1,Y1,...,XN,YN]
constrain(Primary,Tertiary,Energy) :-
        length(Primary,N),
        M is 2*N, M1 is M - 1,
        length(Tertiary,M),
        domain(Tertiary,1,M1),
        starting_point(Tertiary,N),
        avoid_loops(Tertiary),
        next_constraints(Tertiary),
        energy_constraint(Primary,Tertiary,Energy).
```

ENCODING THE PF, HP, ON $\mathbb{N}^2$


```
%%% X1=Y1=X2=N, Y2=N+1
starting_point([N,N,N,N1|_],N) :-
      N1 is N + 1.
avoid_loops(Tertiary):-
       positions_to_integers(Tertiary, ListaInteri),
       all_different(ListaInteri).
positions_to_integers([X,Y|R], [I|S]):-
       I  #= X*100+Y,  %%% 100 is "large"
       positions_to_integers(R,S).
positions_to_integers([],[]).
```

Observe that we do not introduce a disjunction

$$X_i \neq X_j \vee Y_i \neq Y_j$$

for each constraint $(X_i, Y_i) \neq (X_j, Y_j)$

ENCODING THE PF, HP, ON $\mathbb{N}^2$


```
next_constraints([_,_]).
next_constraints([X1,Y1,X2,Y2|C]) :-
        next(X1,Y1,X2,Y2),
        next_constraints([X2,Y2|C]).

next(X1,Y1,X2,Y2):-
        domain([Dx,Dy],0,1),
        Dx #= abs(X1-X2),
        Dy #= abs(Y1-Y2),
        Dx + Dy #= 1.
```

Note: a non linear constraint.

Encoding the PF, HP, on $\mathbb{N}^2$

`energy_constraint(Primary,Tertiary,Energy):-` ...

is defined recursively so as to fix

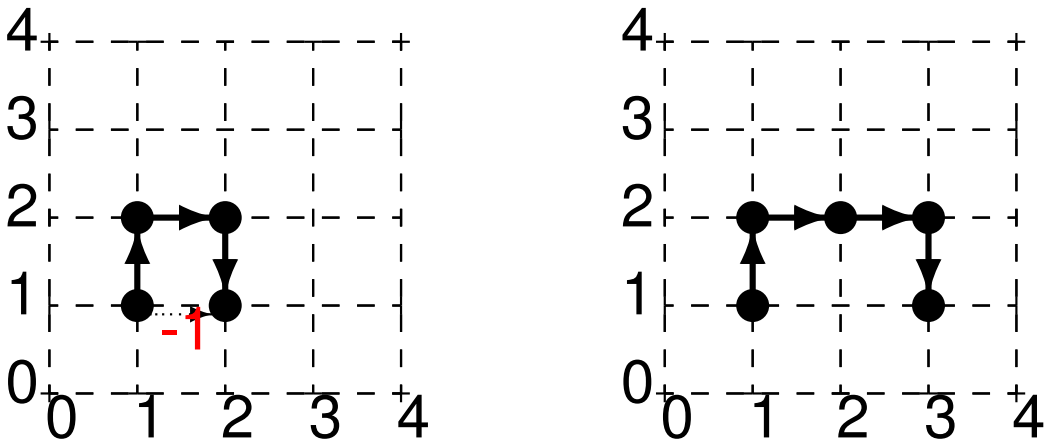
$$\texttt{Energy \#=} C_{1,3} + C_{1,4} + \cdots + C_{1,N} + C_{2,4} + \cdots + C_{2,N} + \vdots + C_{N-2,N}$$

Where each $C_{A,B}$ is defined as follows:

```
energy(h,XA,YA,h,XB,YB,C_AB) :-
    C_AB in {0,-1},
    DX #= abs(XA - XB), DY #= abs(YA - YB),
    1 #= DX + DY  #<=> C_AB #= -1.
energy(h,_,_,p,_,_,0).
energy(p,_,_,h,_,_,0).
energy(p,_,_,p,_,_,0).
```

Constraint Optimization

- This basic code is a good starting point for Optimization.
- A first idea concerns the objective function `Energy`.
- Only aminoacids at an *odd* relative distance can contribute to the Energy.



- Proof: think to the offsets at each step.

Constraint Optimization

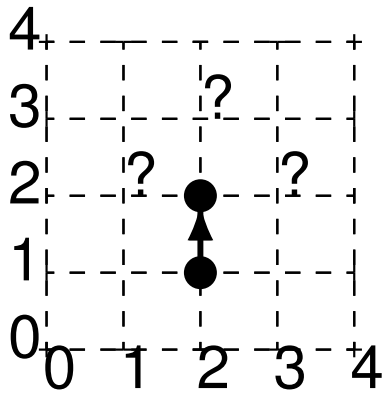
- Thus,

$$\texttt{Energy } \#= \underbrace{C_{1,3}}_{=0} + C_{1,4} + \underbrace{C_{2,4}}_{=0} + \cdots + C_{1,N} +$$
$$\underbrace{C_{2,4}}_{=0} + C_{2,5} + \cdots + C_{2,N} +$$
$$\vdots$$
$$+ \underbrace{C_{N-2,N}}_{=0}$$

- Speed up $3\times$.

Constraint Optimization

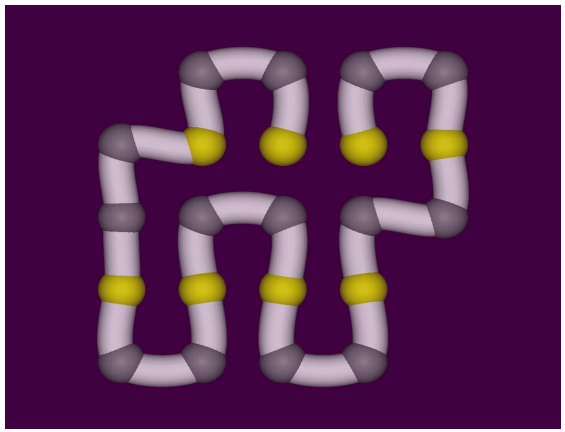
- The next element can (at the beginning) take 3 positions.
- Space looks as $\sim 3^n$
- Precisely, it is $\sim 1.28\frac{2.64^n}{n^{0.34}}$.
- We can reduce it by reducing the domains and/or adding linear distance constraints on pairs of aminoacids.

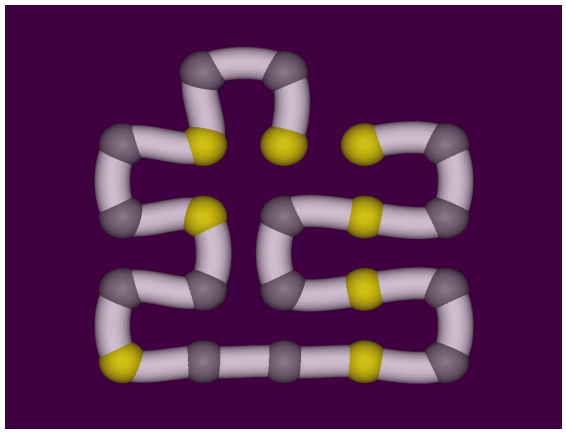


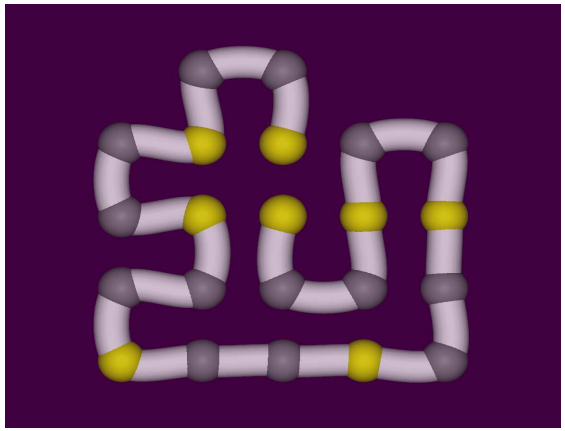
- Typically, in the solutions, the offsets are of the order of $2\sqrt{N}$ (for real proteins there are some more precise formulae).
- Speed up $20\times$.
- Further speed up? Avoid Symmetries (easy in this case). Ad hoc constraint propagation (Backofen–Will).

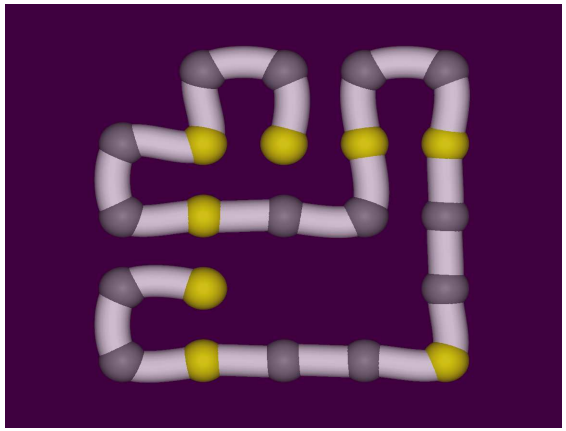
Example

PF([H,P,P,H,P,P,...,H,P,P,H],L)



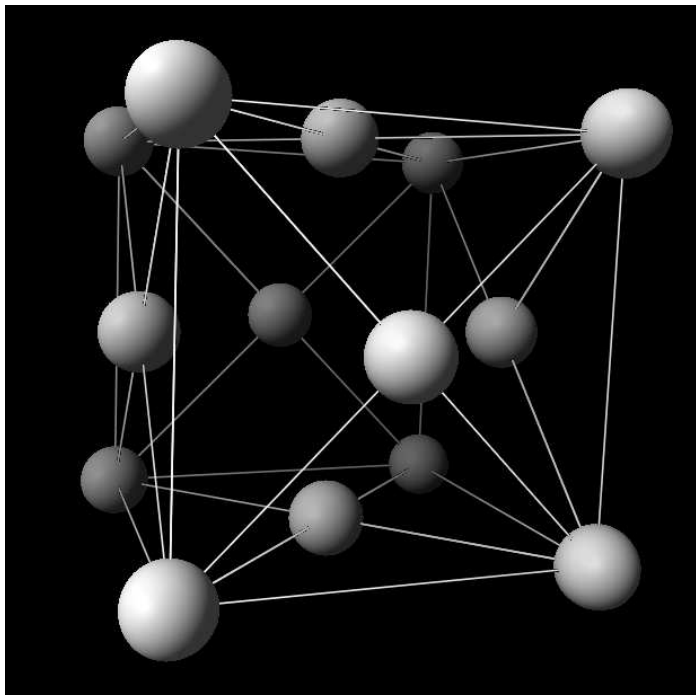






A MORE REALISTIC SPACE MODEL

- The Face Centered Cube lattice models the discrete space in which the protein can fold.
- It is proved to allow realistic conformations.
- The cube has size 2.
- 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 and $60°, 90°, 120°$ and $180°$ bend angles are allowed (in nature $60°$ and $180°$ never occur).



HP ON FCC

- Backofen and Will fold HP-proteins up to length 200
- Clever propagation, an idea of stratification and some geometrical results on the lattice.
- Drawbacks: It is only an abstraction. The solutions obtained are far from reality. For instance, helices and sheets are never obtained.
- Problems:
  - Energy function too simple.
  - Contact too strict.

A more realistic Energy function

- Same assumption: only pairs of aminoacids in *contact* contribute to the energy value.
- The notion of *contact* is easy on lattice models.
- There is a 20x20 *potential matrix* storing the contribution for each pair of aminoacids. [Fogolari et al.]
- Values are either positive or negative.

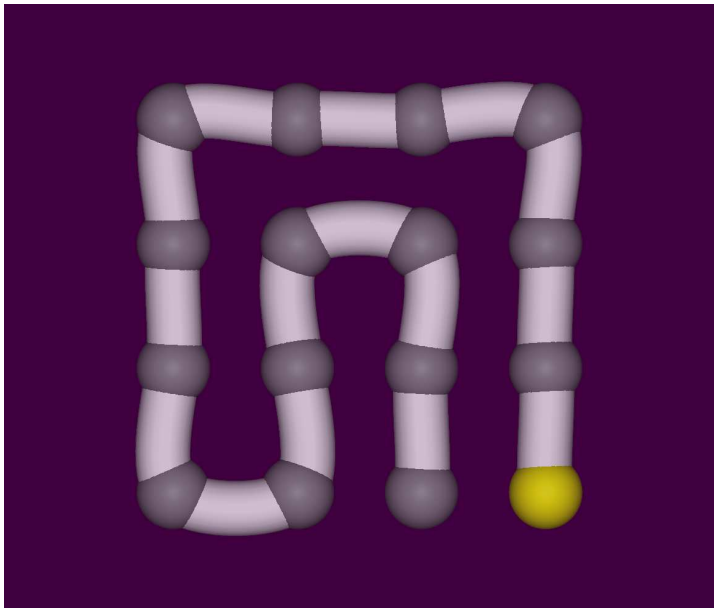
PF, 20 AMINOACIDS, ON $\mathbb{N}^2$

- Basically, the same code, with a call to `table(A,B,Cost).`

```
energy(A,XA,YA,B,XB,YB,C) :-
        table(A,B,Cost),
        (Cost #\= 0,!,
        C in {0,Cost},
        DX #= abs(XA - XB),
        DY #= abs(YA - YB),
        1 #= DX + DY  #<=> C #= Cost;
        C #= 0).
```

Statistical Energy Function
Example

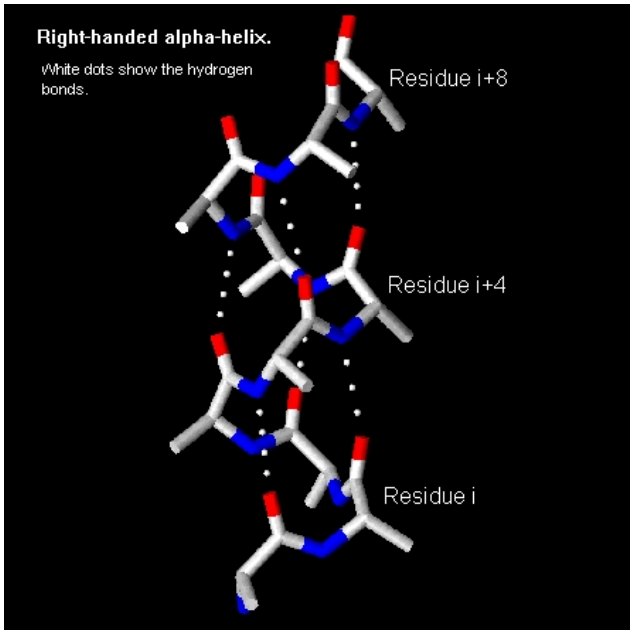
```
pf([s,e,d,g,d,l,p,i,v,a,s,f,m,r,r,d],L).
```
< 1 s, Energy = -7.4. (search space: 6.416.596)



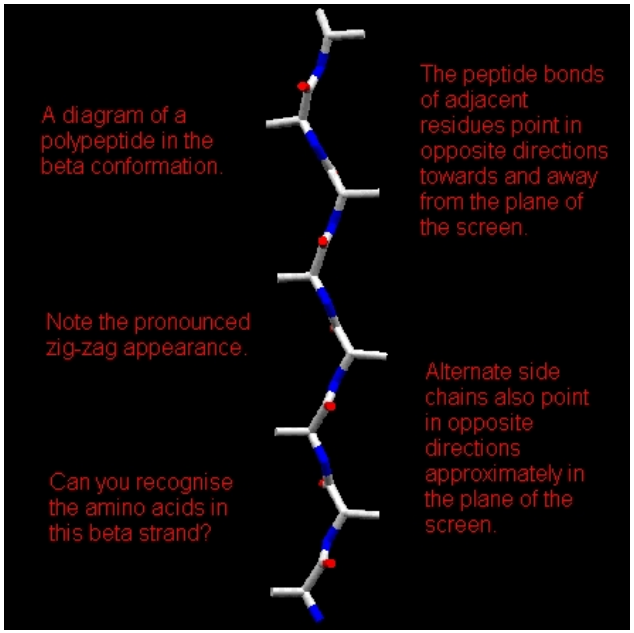
PF, 20 aminoacids, on FCC

- The solution's space is huge $\sim 1.26n^{0.162}(10,03)^n$.
- The potential table is 20×20.
- Contact is set when $|X_1 - X_2| + |Y_1 - Y_2| + |Z_1 - Z_2| = 2$
- New constraints (secondary structure) are needed.
- A careful treatment of the energy function as a matrix that statically (e.g. if two aminoacids s_i, s_j belong to the same α-helix, then $M[i,j] = 0$) and dynamically set to 0 most of its elements.
- Some heuristics for pruning the search tree (loosing completeness!)

Secondary Structure Information



Alpha Helix

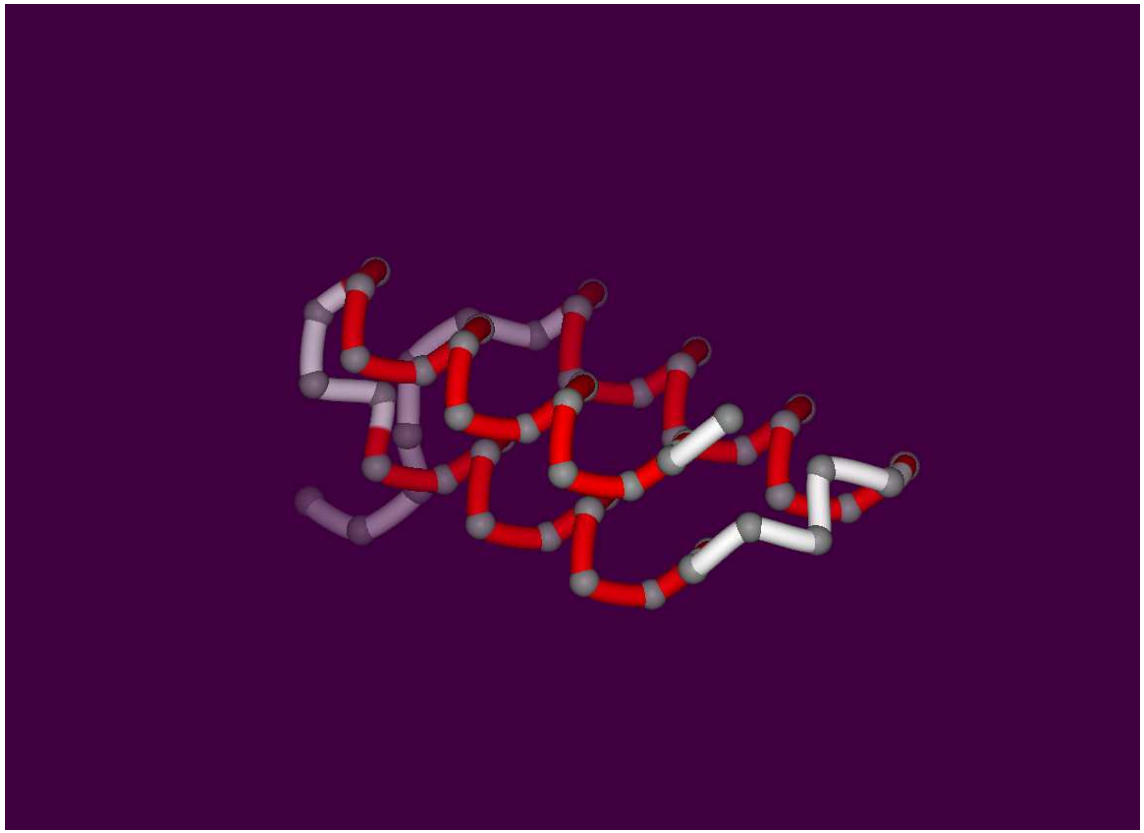


Beta Sheet

- Secondary structure can be predicted
- Distance constraints are set to force these substructures
- ssbonds constraints induced by aminoacids: **Cys**teine ($C_3H_7NO_2S$) and **Met**hionine ($C_5H_{11}NO_2S$) are set as well.

EXAMPLE


```
protein('1ENH', Primary, Secondary):-
   Primary = [r,p,r,t,a,f,s,s,e,q,
   l,a,r,l,k,r,e,f,n,e, n,r,y,l,t,e,r,r,r,q,
   q,l,s,s,e,l,g,l,n,e, a,q,i,k,i,w,f,q,n,k, r,a,k,i],
   Secondary = [helix(8,20),helix(26,36),
                helix(40,52), strand(22,23)].
```



Evoluzione della Codifica 20x20 su FCC

- Primo encoding in SICStus: *Dovier, Fogolari, Burato, Grado–WFLP02, ENTCS76*
- Evoluzione con miglior trattamento struttura secondaria e matrice dinamica: *Dal Palù, Dovier, Fogolari, Budapest–ERCIM03 LNAI 3010, BMC Bioinformatics 2004*
- Riorganizzazione codice, sviluppo di euristiche di ricerca (BBF) e parallelismo: *Dal Palù, Dovier, Pontelli Lisbona–PPDP 2005, ACM*

Evoluzione della Codifica 20x20 su FCC

- Sviluppo e miglioramento del risolutore ad-hoc COLA (COnstraint solving on LAttices: *Dal Palù, Dovier, Pontelli Montego Bay-LPAR 2005—SPE 2007*
- Lavori più generali sui *vincoli globali* per il problema (Nantes–WCB06, Porto–WCB07, La Jolla–BIMB07, J. DMB08) e sull'uso di tecniche di model checking per affrontare lo spazio di ricerca.
- (in corso) Inserimento efficiente vincoli globali in COLA, incluse mappe di densità, posizionamento catene laterali, tecniche ibride constraint+local search.