

---

# *MPRI C2-19*

## **Protein Structure Prediction by Constraint Logic Programming**

François Fages,  
Constraint Programming Group,  
INRIA Rocquencourt

`mailto:Francois.Fages@inria.fr`  
`http://contraintes.inria.fr/`



# Molecules in the Cell

Small molecules: covalent bonds 50-200 kcal/mol

- " 70% water, 1% ions, 6% amino acids (20), nucleotides (5),
- " fats, sugars, ATP, ADP, &

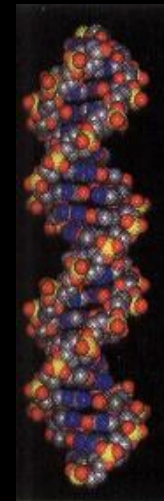
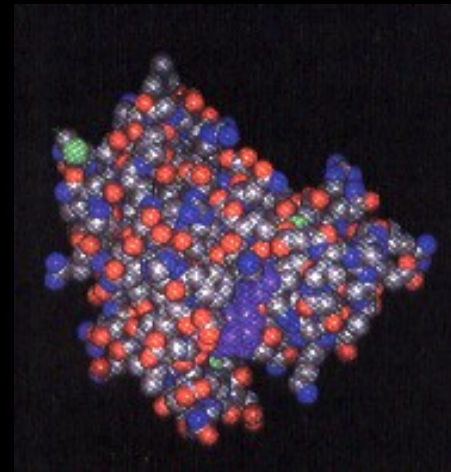
Macromolecules: hydrogen bonds, ionic, hydrophobic, Waals 1-5 kcal/mol

*Stability and bindings determined by the number of weak bonds: 3D shape*

" 20% proteins (50-10<sup>4</sup> amino acids)

" RNA (10<sup>2</sup>-10<sup>4</sup> nucleotides AGCU)

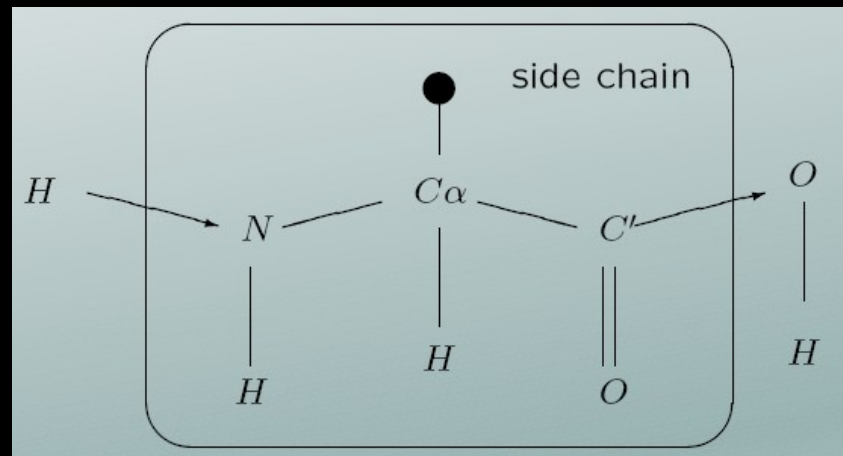
" DNA (10<sup>2</sup>-10<sup>6</sup> nucleotides AGCT)



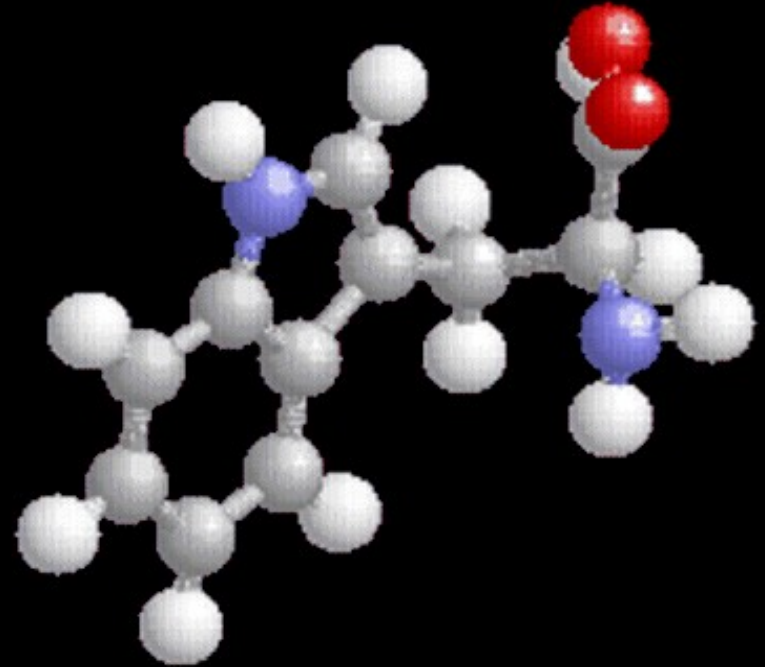
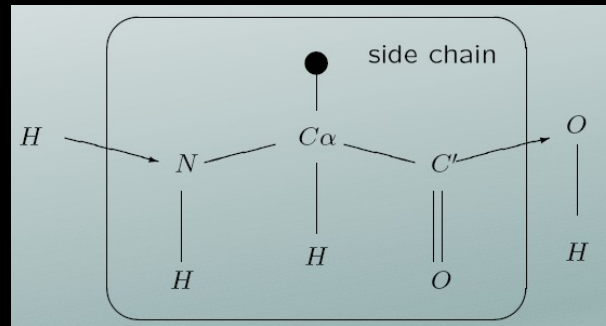
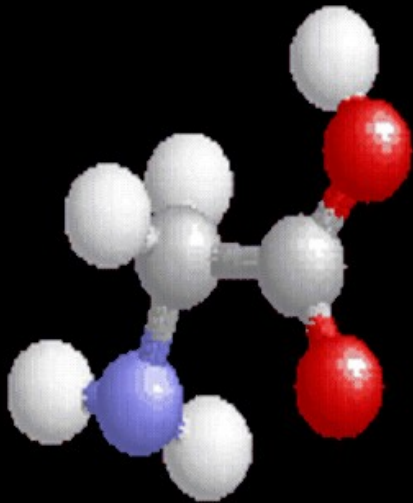
# Aminoacids

20 aminoacids: Alanine (A), Cysteine (C), Aspartic Acid (D), Glutamic Acid (E), Phenylalanine (F), Glycine (G), Histidine (H), Isoleucine (I), Lysine (K), Leucine (L), Methionine (M), Asparagine (N), Proline (P), Glutamine (Q), Arginine (R), Serine (S), Threonine (T), Valine (V), Tryptophan (W), Tyrosine (Y).

Same backbone centered on  $C\alpha$  linked with covalent C-N bonds to different side chains (residues) from 1 atom (for G) to 18 (for T)



# Examples: Glycine and Tryptophan



$C_2H_5NO_2 \rightarrow 9+1=10$  atoms

$C_{11}H_{12}N_2O_2 \rightarrow 9+18=27$  atoms

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

<http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids.en.html>

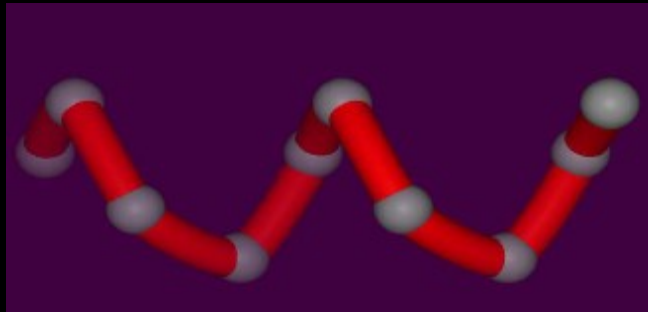


# Secondary Structure of Proteins

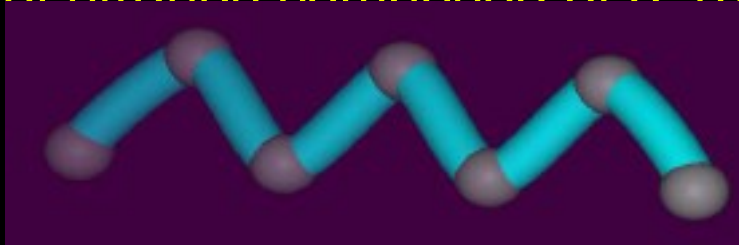
Protein = word of  $m$  forms

among three forms ( $3^m$  possibilities) stabilized by *hydrogen bonds*  $H\cdots O$ :

∇  $\alpha$ -helices of 5-40 contiguous residues (with 3.6 res. per tour)



∇  $\beta$ -sheets of strands composed of 5-10 contiguous residues

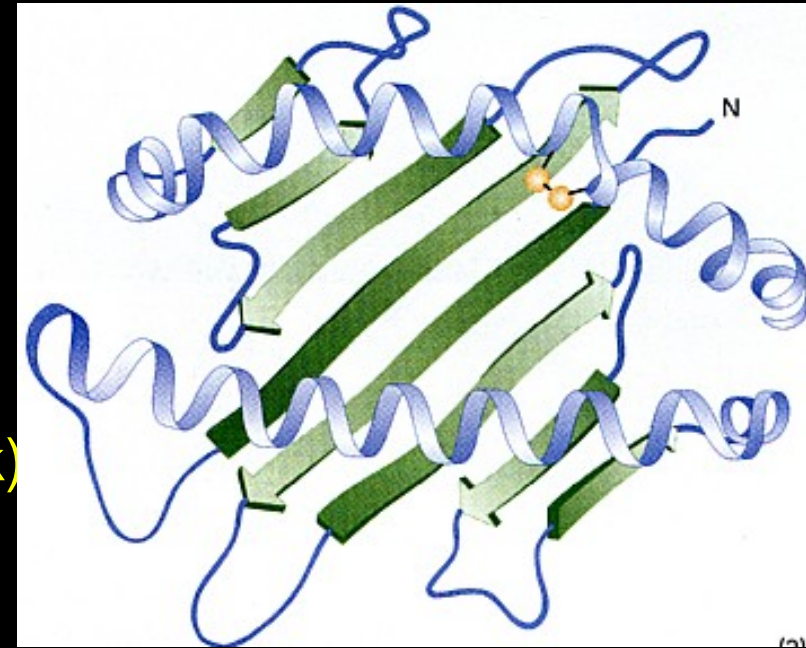


" random coils,&

# Tertiary Structure of Proteins

Protein= 3D spatial conformation

- " Native conformation in a determined environment (e.g. water)
- " 30000 structures in Protein Data Bank
- " Bonded interactions between atoms:
  - covalent bonds (unbreakable)
  - disulfide bonds (slow to form/break)  
e.g. between cysteine residues
  - hydrogen bonds H O (fast to form/break)
- " Non-bonded interactions between atoms:
  - electrostatic (long-range)
  - van der Waals (short range)
- " Confirmation that minimizes the free energy (Anfinsen's hypothesis)



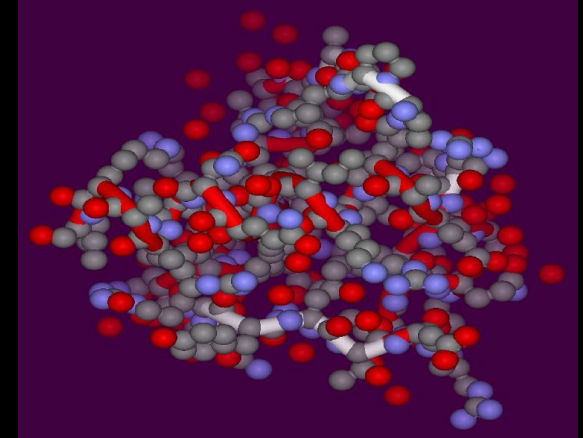
# Importance of Protein Structure Prediction

- " Understand protein folding, interaction capabilities, protein docking
- " Domain prediction, function prediction
- " Drug design and/or optimization
  - More than 50% of the drugs target receptor proteins
- " Enzymes design and/or optimization
- " Inverse problem: protein synthesis of a given shape
  - Can restrict the number of amino acids
  
- " One of the most important problems of bioinformatics
- " Methods are evaluated in the CASP competition (every two years)
- " Protein data bank: repository of tertiary structures (weekly updates)

# Protein Structure Prediction and Folding Problems

Input: the primary structure of a protein i.e. sequence of  $n$  amino acids (plus some information on its secondary structure)

Output (PSP): the tertiary structure of the protein that minimizes the free energy.



Output (PFP): the folding sequence to the tertiary structure of the protein (with secondary structures formed first)

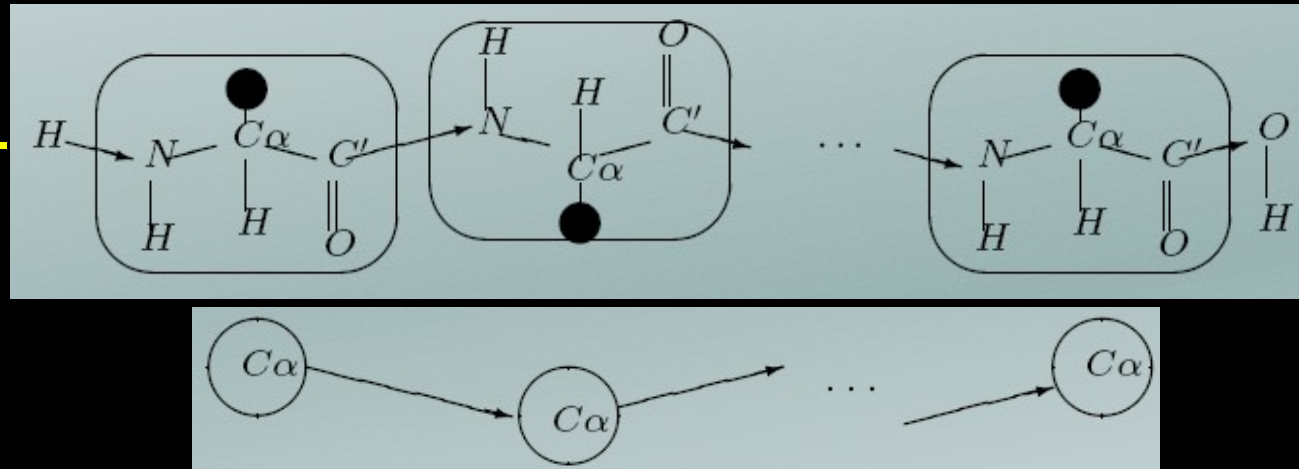
*E.g. Hypotheses in the simple HP model:*

- " Each aminoacid is a sphere centered on its  $C_{\alpha}$  atom
- " Each aminoacid is either hydrophobic H or polar P (hydrophilic)
- "  $\text{Energy(HH)} = -1$   $\text{Energy(HP,PH,PP)} = 0$

# Protein Structure Prediction Problem

" Spatial model: in known protein structures, the distance between two consecutive  $C\alpha$  atoms is essentially fixed (3.8° A)

- Lattice (discrete) models.
- Continuous models.



2) Energy model:

- HP model:  $\text{Energy}(\text{HH}) = -1$   $\text{Energy}(\text{HP}, \text{PH}, \text{PP}) = 0$
- 20x20 Matrix model:  $\text{Energy}(\text{AB})$

# HP Energy Model

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

Energy(HH)= -1 Energy(HP,PH,PP)=0 for pairs of aminoacids in contact

# Matrix Energy Model

- " Same assumption: only pairs of aminoacids in contact contribute to the energy value.
- " There is a potential matrix storing the contribution for each pair of aminoacids in contact.
- " Values are either positive or negative.
- " The global energy must be minimized.

Energy(AB)= $w_{AB}$  for pairs on aminoacids in contact



# Admissible Foldings

Let  $S = s_1, \dots, s_n$  be a sequence of amino acids

Def. A folding in a crystal lattice  $(P, E)$  is a mapping  $w: N \rightarrow P$ ,  $w(i) = P_i$



# Admissible Foldings

Let  $S=s_1, \dots, s_n$  be a sequence of amino acids

Def. A folding in a crystal lattice  $(P, E)$  is a mapping  $w: N \rightarrow P$ ,  $w(i) = P_i$

Def. An admissible folding is a folding such that

- " Constant distance  $k$  for consecutive pairs:  $\text{eucl}(w(i), w(i+1)) = k$
- " Non-overlapping:  $\text{eucl}(w(i), w(j)) > 0$  for  $i \neq j$
- " Occupancy of side chain:  $\text{eucl}(w(i), w(j)) > k$  for  $|i-j| \geq 2$
- " Bend angles in  $[90^\circ .. 150^\circ]$ :  $k_1 \leq \text{eucl}(w(i), w(i+2)) \leq k_2$  for  $i \in \{1, \dots, n-2\}$

---

# Crystal Lattice Models

Def. A crystal lattice is a graph  $(P, E)$  where  $P$  is a set of points in  $\mathbb{Z}^3$  connected by undirected edges  $E$ .



# Crystal Lattice Models

Def. A crystal lattice is a graph  $(P,E)$  where  $P$  is a set of points in  $\mathbb{Z}^3$  connected by undirected edges  $E$ .

Def. Squared euclidean distance  $eucl(A,B)=(x_B-x_A)^2+(y_B-y_A)^2+(z_B-z_A)^2$

# Crystal Lattice Models

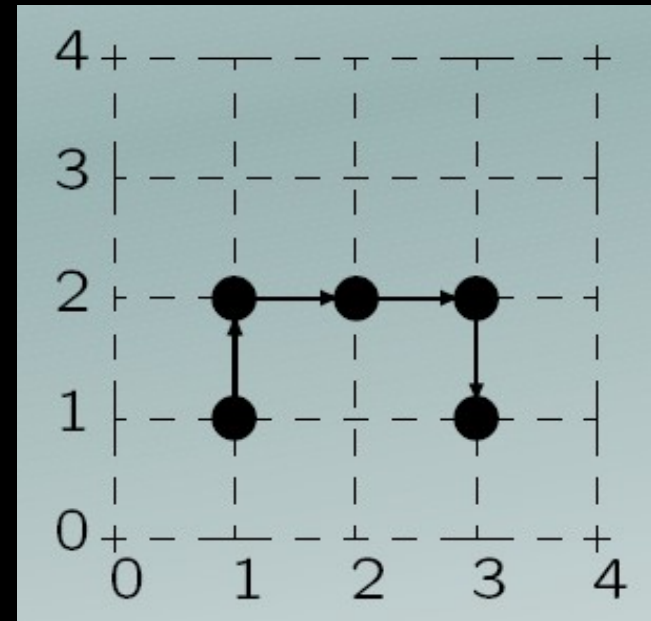
Def. A crystal lattice is a graph  $(P,E)$  where  $P$  is a set of points in  $\mathbb{Z}^3$  connected by undirected edges  $E$ .

Def. Squared euclidean distance  $eucl(A,B)=(x_B-x_A)^2+(y_B-y_A)^2+(z_B-z_A)^2$

Def. A cubic lattice is a crystal lattice  $(P,E)$  such that

- "  $P=\{(x,y,z) \mid x,y,z \in \mathbb{Z}\}$
- "  $E=\{(A,B) \mid A,B \in P, eucl(A,B)=k=1\}$

A cubic lattice is 6-connected.  
(solve  $x^2+y^2+z^2=1$  in  $(0,0,0)$ )



# PSP in a Cubic Lattice Model

- " Cubes of size 1,
- " Vertices labeled by aminoacid names

Find a conformation  $w:\{1 \dots n\} \rightarrow \mathbb{Z}^3$  such that:

- "  $w(i) \neq w(j)$  for  $i \neq j$
- "  $\|w(i) - w(i+1)\| = 1$  for the Euclidean norm
- " Minimizes the energy  $E = \sum_{\|w(i) - w(j)\| = 1, i+1 < j} \text{Energy}(w(i), w(j))$

Thm. The existence of an admissible conformation with energy less than a constant  $k$  is NP-complete.

Proof. Reduce the bin packing problem to a PSP problem

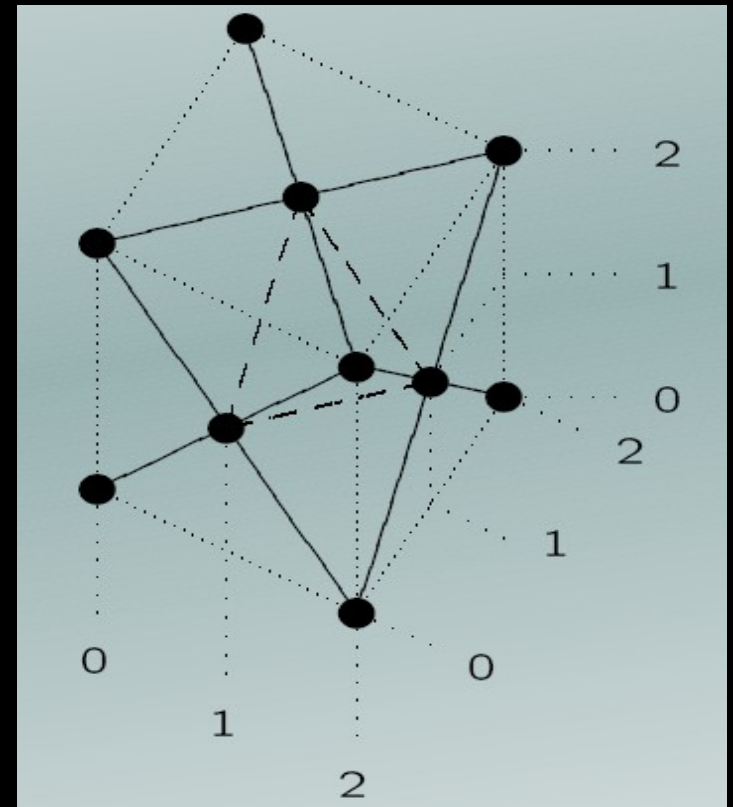


# Face-Centered Cubic Lattice

Def. A face-centered cubic lattice is a crystal lattice  $(P,E)$  such that

- "  $P = \{(x,y,z) \mid x,y,z \in \mathbb{Z}, x+y+z \text{ is even}\}$
- "  $E = \{(A,B) \mid A,B \in P, \text{eucl}(A,B) = \sqrt{2}\}$

An FCC is 6-connected at distance 2  
(solve  $x^2+y^2+z^2=4$  in  $(0,0,0)$ )



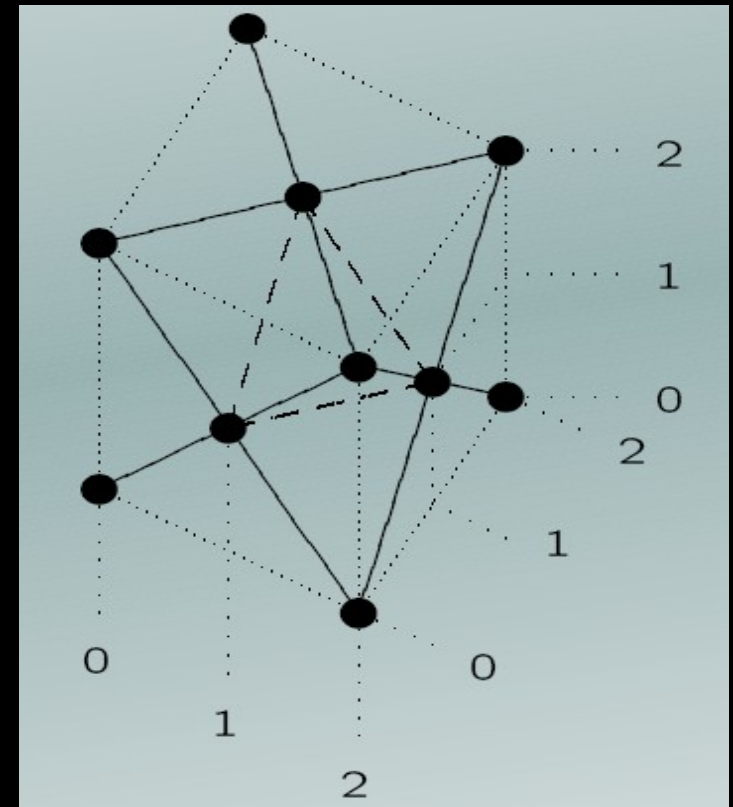
# Face-Centered Cubic Lattice

Def. A face-centered cubic lattice is a crystal lattice  $(P,E)$  such that

- "  $P = \{(x,y,z) \mid x,y,z \in \mathbb{Z}, x+y+z \text{ is even}\}$
- "  $E = \{(A,B) \mid A,B \in P, \text{eucl}(A,B) = \sqrt{2}\}$

An FCC is 6-connected at distance 2

?-connected at distance  $\sqrt{2}$  ?



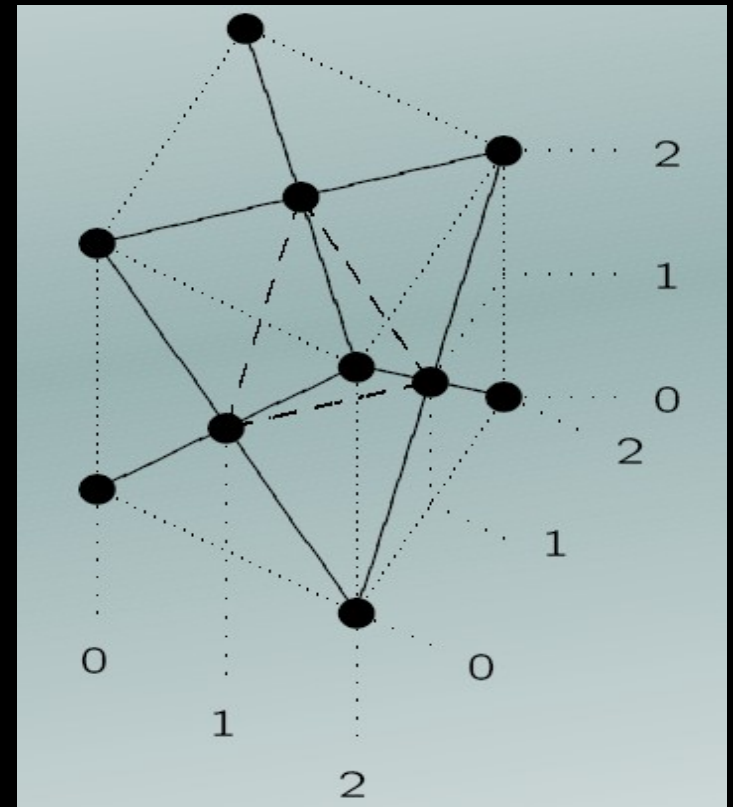
# Face-Centered Cubic Lattice

Def. A face-centered cubic lattice is a crystal lattice  $(P,E)$  such that

- "  $P = \{(x,y,z) \mid x,y,z \in \mathbb{Z}, x+y+z \text{ is even}\}$
- "  $E = \{(A,B) \mid A,B \in P, \text{eucl}(A,B) = \sqrt{2}\}$

An FCC is 6-connected at distance 2  
12-connected at distance  $\sqrt{2}$   
(solve  $x^2+y^2+z^2=2$  in  $(0,0,0)$ )  
18 contact neighbours at distance  $\leq 2$

? bend angles ?



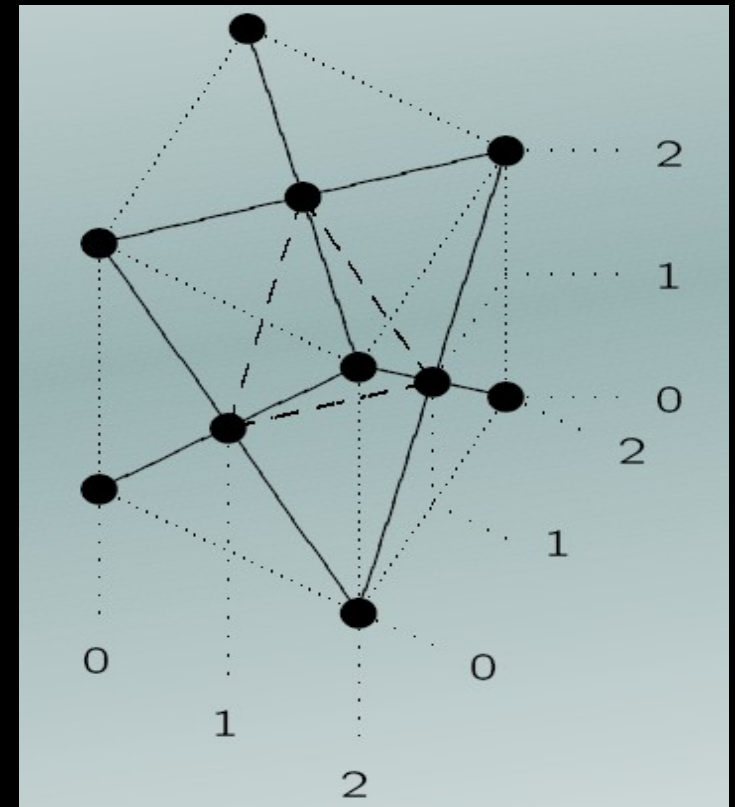
# Face-Centered Cubic Lattice

Def. A face-centered cubic lattice is a crystal lattice  $(P,E)$  such that

- "  $P = \{(x,y,z) \mid x,y,z \in \mathbb{Z}, x+y+z \text{ is even}\}$
- "  $E = \{(A,B) \mid A,B \in P, \text{eucl}(A,B) = \sqrt{2}\}$

An FCC is 6-connected at distance 2  
12-connected at distance  $\sqrt{2}$   
18 contact neighbours at distance  $\leq 2$

$(60^\circ)$ ,  $90^\circ$ ,  $120^\circ$  and  $180^\circ$  bend angles



# Face-Centered Cubic Lattice Model

## *Abstraction:*

- " Each aminoacid is a sphere centered on its  $C_{\alpha}$  atom
- " Fixed distance of 3.8Å between  $C_{\alpha}$  atoms
- " Table  $Pot(x,y)$  of energy value per pairs of aminoacids  $x$  and  $y$

## *Face-centered Cubic Lattice model:*

- " Cubes of size 2 where points are *vertices and central points of faces*
- " 12 connected neighbors at distance  $\sqrt{2}$  (corresponding to 3.8Å)
- " 18 contact neighbors at distance  $\leq 2$  (corresponding to 5.4Å < 6.4Å)

## *Conformation $w: \{1 \& n\} \rightarrow \mathbb{Z}^3$ such that:*

- "  $w(i) \neq w(j)$  for  $i \neq j$  and  $\|w(i) - w(i+1)\| = \sqrt{2}$
- " Minimizes the energy  $E = \sum_{\|w(i) - w(j)\| = 2, i+1 < j} Pot(w(i), w(j))$

# Principles of Constraint Logic Programming

## " Constrain-and-generate versus generate-and-test.

```
find(X):-
    constrain(X),
    generate(X).

find(X):-
    generate(X),
    test(X).
```

active use of constraints to prune the search tree in advance

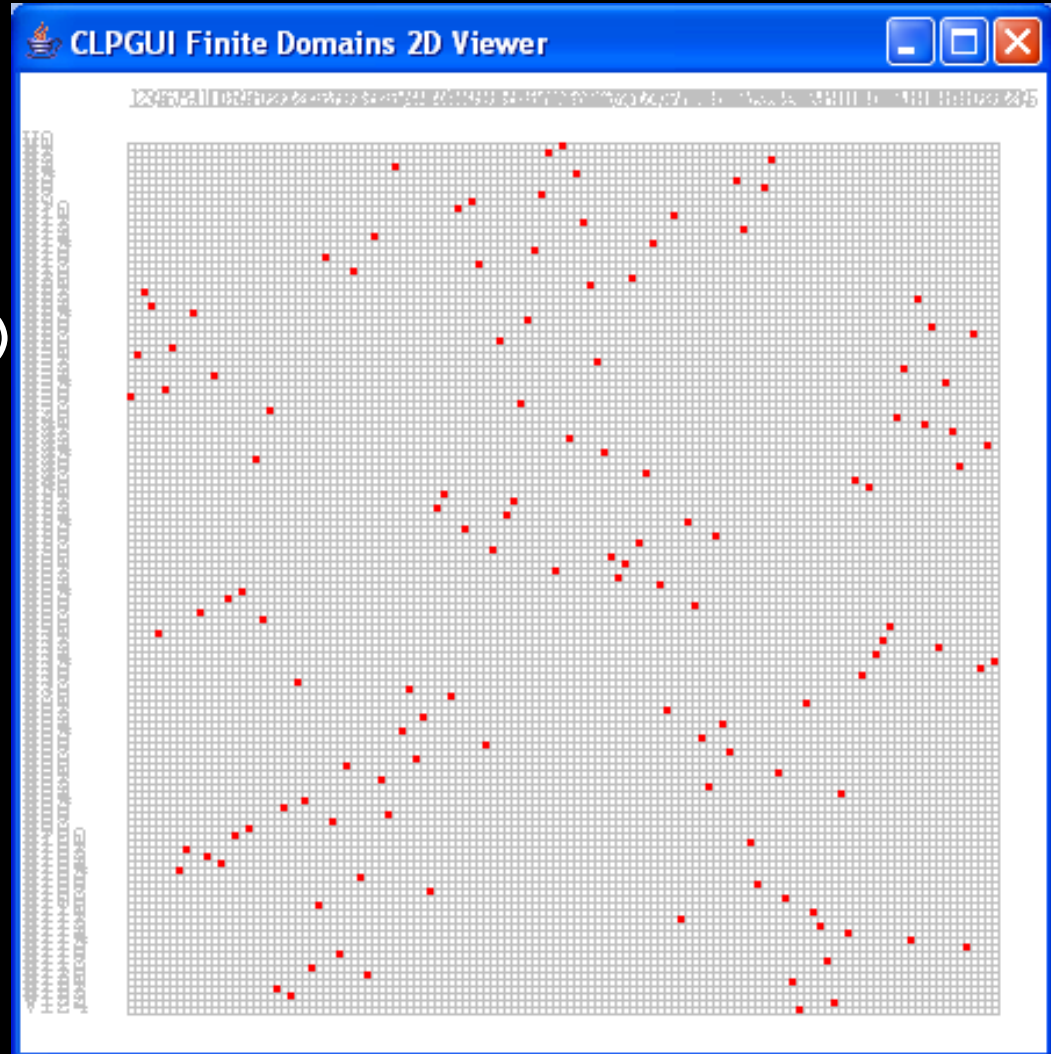
- symbolic/numerical solving of constraints
- constraint propagation over finite domains

## " Optimization by branch-and-bound procedure

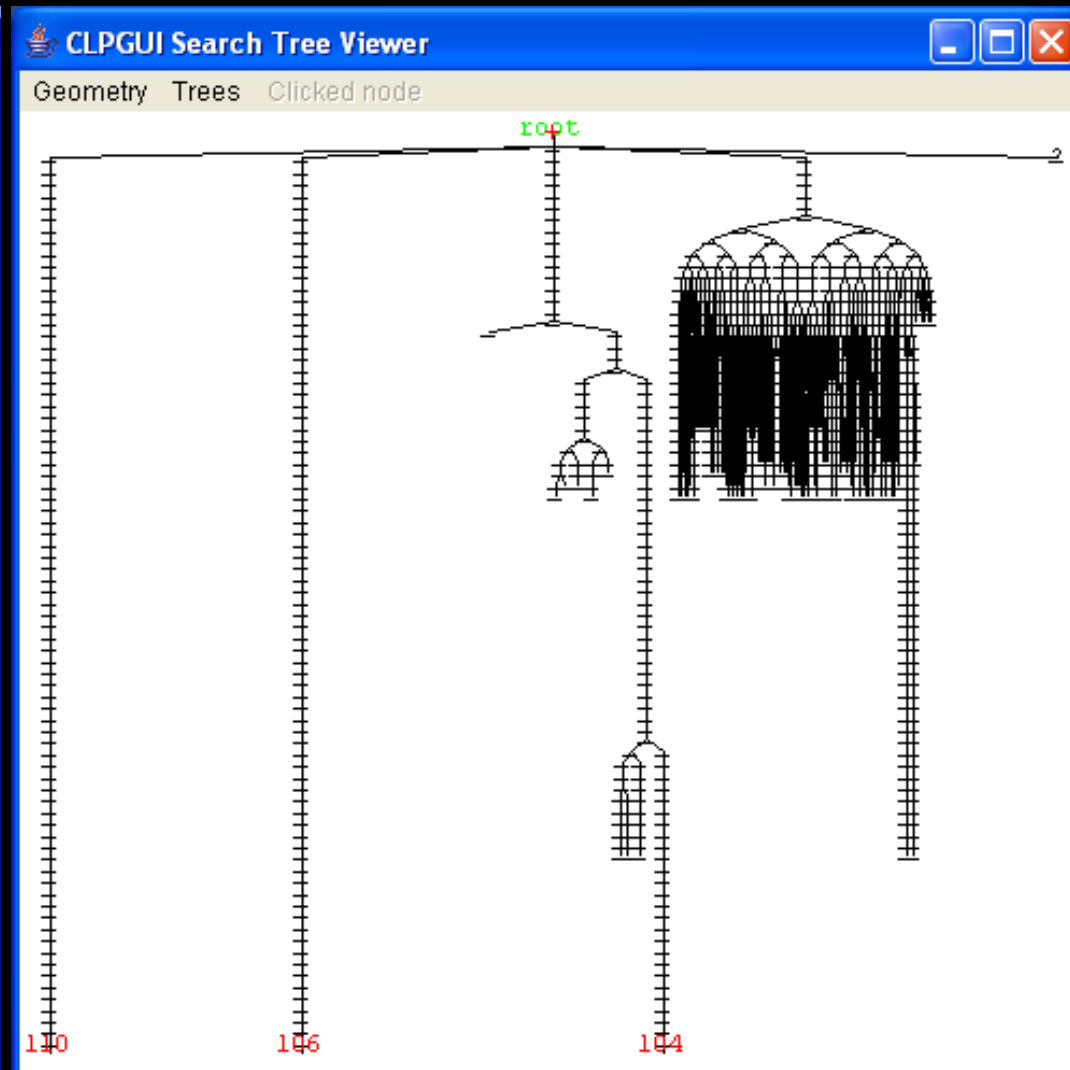
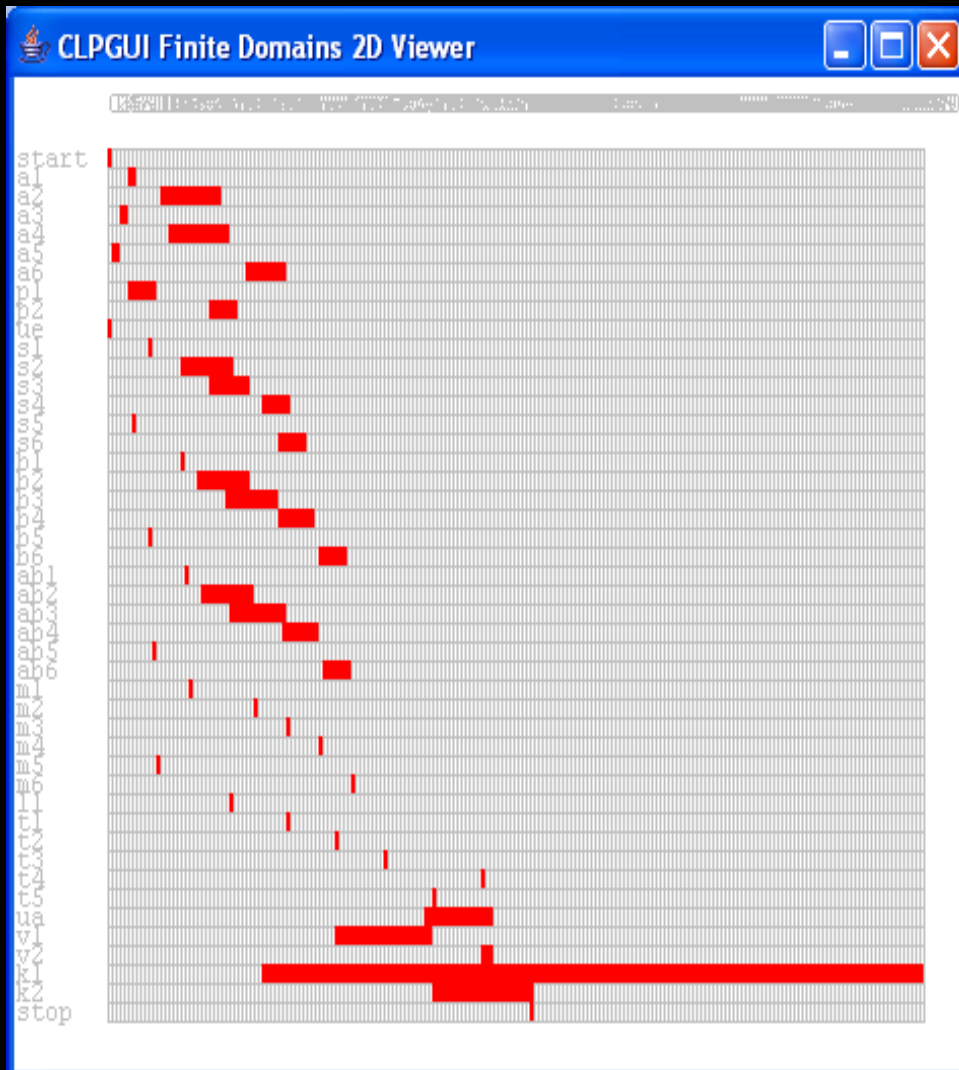
```
optim(X,C,R):- cost(X)<C, find(X), optim(Y,cost(X),R).
optim(X,C,C).
```

# Placing N queens on an NxN Chessboard

```
queens (N, L) :-  
    length(L, N),  
    fd_domain(L, 1, N),  
    safe(L),  
    fd_labeling(L, first_fail)  
safe([]).  
safe([X|L]) :-  
    noattack(L, X, 1),  
    safe(L).  
noattack([], _, _).  
noattack([Y|L], X, I) :-  
    X#\=Y,  
    X#\=Y+I,  
    X+I#\=Y,  
    I1 is I+1,  
    noattack(L, X, I1).
```



# Optimal Scheduling of a Project of 50 Tasks



# CLP Approaches to Protein Structure Predication

- " HP cubic model [Backofen-Will 03]
- " HP face-centered cubic model [Dal Palu-Dovier-Fogolari04]  
from seconds for  $n < 50$  to tenth of minutes for  $n < 100$

```
fcc_pf( ID, Time, Compact) :-  
    protein(ID, Primary, Secondary) ,  
    constrain(Primary, Secondary, Indexes ,  
              Tertiary, Energy, Matrix, Freq, Compact) ,  
    writetime(video, 'Constraint time: '), nl, ! ,  
    solution_search(Time, Primary, Secondary, Indexes ,  
                   Tertiary, Energy, Matrix, Freq) ,  
  
    print_results(ID, Time, Primary, Secondary, Tertiary, Compact) .
```

# Constraint Logic Program [Dal Palu-Dovier-Fogolari04]

## " Domain bounds

```
domain_bounds(N, [X,Y,Z|Rest]):- CUBESIZE is N * 2, domain([X,Y,Z],0,CUBESIZE),
                                sum([X,Y,Z],#=#,Sum), even(Sum), domain_bounds(N, Rest).
domain_bounds(_, []).
```

## " Circuit constraint

```
avoid_self_loops(Tertiary, N):- BASE is 2*N+1, positions_to_integers(Tertiary,ListIntegers,BASE),
                                all_different(ListIntegers).
```

## " Connectivity constraints

```
next(X1,Y1,Z1,X2,Y2,Z2):- domain([Dx,Dy,Dz],0,1),
                            Dx #= abs(X1-X2), Dy #= abs(Y1-Y2), Dz #= abs(Z1-Z2),
                            sum([Dx,Dy,Dz], #=, 2).
```

## " Non-connectivity constraints

```
non_next(X1,Y1,Z1,X2,Y2,Z2):- Dx#=(X1-X2)*(X1-X2), Dy#=(Y1-Y2)*(Y1-Y2),
                                Dz#=(Z1-Z2)*(Z1-Z2), sum([Dx,Dy,Dz], #>, 2).
```

## " Contact energy constraint

# Explored Search Tree for Protein 1KVG (n=12)

Size	Time
17	0,04
27	1,76
34	0,8
36	4,31
6	45,3
63	58mn
8	2mn
9	1,18
97	35mn
104	1 0mn

