Protein Structure Prediction

Ulf Nilsson, IDA

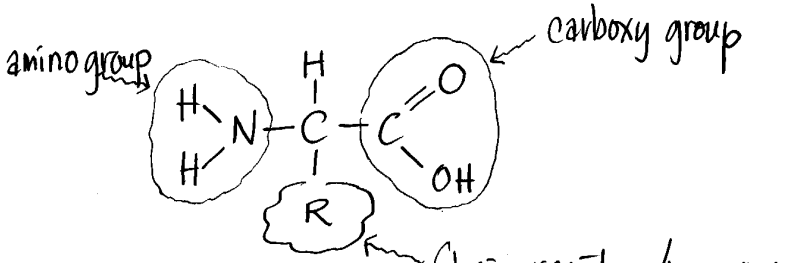
2001-11-23

Outline

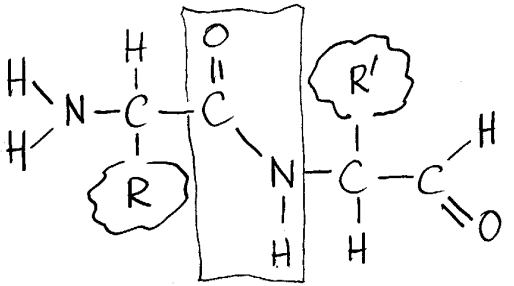
- Problem description
- The HP-model
- A naive algorithm
- Monte Carlo simulations
- Constraint based approach
  - Constraints
  - Basic notions
  - Finite domain constraints
  - Propagation and search
  - A constraint encoding of the HP-model.

Amino acids

There are 20 amino acids



Protein (sequence of amino acids)



Protein Structure Prediction

Each (natural) protein folds into a unique three-dimensional structure (conformation) determined by its sequence of amino acids.

The function of the protein is determined by its native conformation.

The protein structure prediction problem

Given the sequence of amino acids, determine the native conformation (tertiary structure).

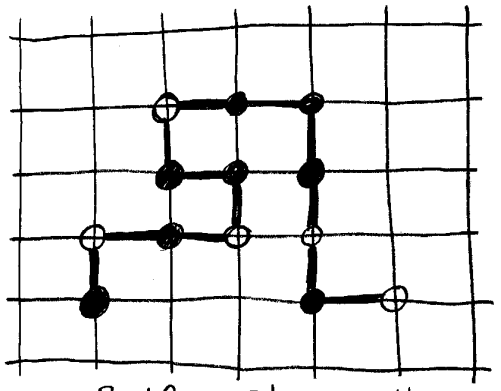
Other, related, problems

- The folding process
- Sequence synthesis (protein engineering)

A Simplified model

- Unisize monomers (residues)
- Unisize bond length
- Monomers are positioned on a lattice
- Simplified energy function
- Two types of monomers [Lau & Dill]
  - Hydrophobic, H
  - Polar, P (hydrophilic)

HPHPHHPHHHPHP



Self-avoiding walk

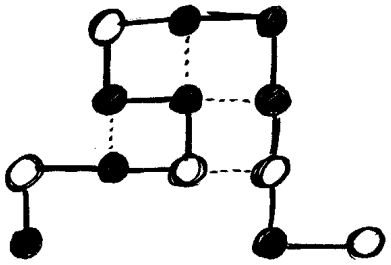
Contact

Two monomers are in <u>contact</u> if they are non-consequtive and distance 1 apart.

Energy

The energy function of the HP-model is given by the following contacts

	H	P
H	-1	0
P	0	0



Energy: 3
Contacts: 4

Basic Definitions

A protein is a sequence $s \in \{H,P\}^*$

If $s = x_1 x_2 \ldots x_n$ then s_i denotes x_i.

Ex $s = HPHHPHPP$

$s_1 = H$, $s_2 = P$, $s_3 = H$ etc.

A monomer s_i is even if i is even.
A monomer s_i is odd if i is odd.

A conformation of a sequence s is a function $c : [1..|s|] \to \mathbb{Z}^d$ where $d \in \{2,3\}$. such that

- $\| c(i) - c(i+1) \| = 1$ for all $1 \le i < |s|$
- $c(i) \neq c(j)$ whenever $i \neq j$.

Monte Carlo w. Metropolis criterion

1. Generate a random conformation
2. Repeatedly attempt local moves until the resulting conformation is self-avoiding
3. Compute the energy of the new conformation. If lower than the previous conformation, accept the move. Otherwise accept the move at random (depending on the energy difference).
4. Record the conformation if the new energy is minimal.

   Repeat 2-4 a large number of times (50.000.000???).

Naive Algorithm

```
INPUT:  A string s[1..n];
OUTPUT: The minimal energy of s;

main
    minE := 0;
    E := 0;
    c[1] := (0,0);
    c[2] := (1,0);
    fold(3, n);
    return minE;

fold(i, j)
    if( i > j ) then   % The whole string is folded
       if new minimum then minE := E
    else
      for all unoccupied neighbors (x,y) of c[i-1] do
        c[i] = (x,y);  % Make a move
        if( s[i] = "H" ) then
           update E with all HH-contacts involving c[i]
        fold(i+1, j);  % Fold the rest of the string
        restore E;     % Undo update to E, before trying next move
      od
    fi
```

Monte Carlo moves

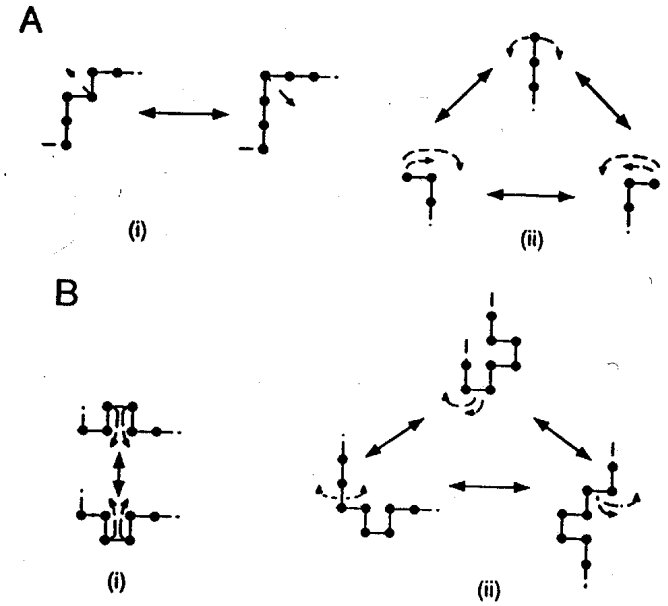


Fig. 41. Move sets used in 2D exact lattice enumeration studies of chain dynamics (Chan & Dill, 1993b, 1994). Double-headed arrows show which conformations are adjacent. Dashed arrows show monomer moves. **A:** Move set 1 (MS1); (i) a three-bead flip; (ii) end flips. **B:** Move set 2 (MS2) includes those in MS1 and also (i) crankshaft moves; (ii) rigid rotations.

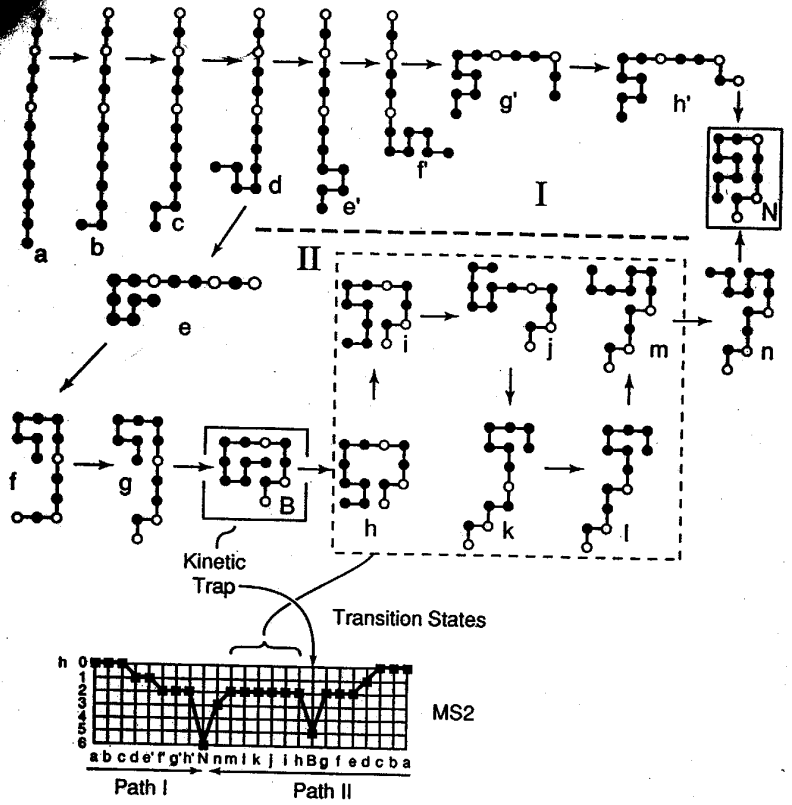


Fig. 43. Typical folding paths and their energy landscapes. Chains begin at conformation "a" and proceed to the native structure N. Path I has no barriers to N (a "throughway" path), but path II passes through local-minimum conformation B, then uphill across transition-state conformations to N. Bottom plot gives the history of the number of HH contacts *h*, and is the energy landscape along the two paths. (From Chan and Dill [1994].) Both paths begin with a hydrophobic zipper collapse.

Constraint

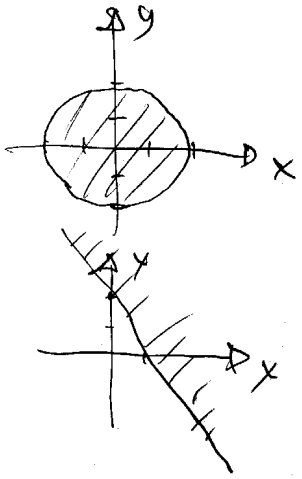
Given a set of variables, a *constraint* is a restriction on the possible values of the variables.

Example

Variables: X, Y.

Constraint I: $X^2 + Y^2 \leq 4$

Constraint II: $Y \geq 2 - 2 \cdot X$



Solution

The constraint $X^2 + Y^2 \leq 4$ has a set of *solutions* – variable assignments when the constraint is true, e.g:

$\{X \mapsto 2, Y \mapsto 0\}$
$\{X \mapsto 0, Y \mapsto 2\}$
$\{X \mapsto 1, Y \mapsto 1\}$

A mapping from variables to values is called a *valuation*. A valuation where the constraint is true is called a *solution*.

Domain of a constraint

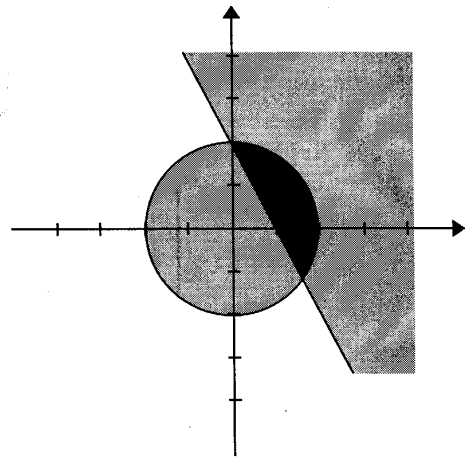
Whether a constraint has a solution or not depends on the values that the variables can take.

The constraint $X^2 = 2$ has a real solution, but not an integer or a rational solution.

The set of all possible values of the variables is called the *domain* of the constraint.

Conjunctive constraints

The conjunction of the primitive constraints $X^2 + Y^2 \leq 4$ and $Y \geq 2 - 2 \cdot X$ is a new (conjunctive) constraint:



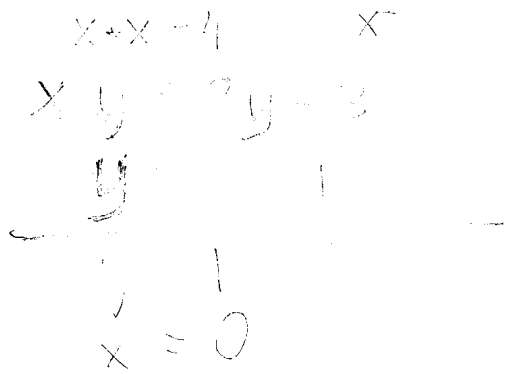
Sets of primitive constraints represent conjunctive constraints.

Properties of constraints

A constraint is said to be *satisfiable* iff it has at least one solution.

A constraint C_1 *implies* a constraint C_2 (written $C_1 \models C_2$) iff every solution of C_1 is also a solution of C_2.

Two constraints are *equivalent* if they have the same set of solutions.



Optimal solutions

A solution σ of a set of constraints S is *maximal subject to* an expression E if $\sigma(E)$ is greater than $\sigma'(E)$ for any solution σ' of S.

Example

The solution $\{X \mapsto 1.6, Y \mapsto -1.2\}$ is a maximal solution of

$$\begin{aligned} X^2 + Y^2 &\leq 4 \\ Y &\geq 2 - 2 \cdot X \end{aligned}$$

subject to $-Y$.

Finite Domain Constraints

Constraints involving variables ranging over finite sets (usually of integers).

Definition of domains

X in 1..8
Y in 1..8
Z in 1..8

Example of constraints:

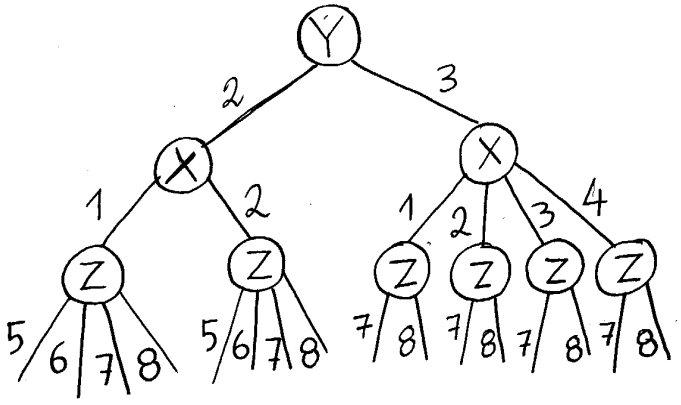
$(X+1 < 2Y) \wedge (2Y < Z)$

Propagation

$X+1 < 2Y \wedge 2Y < Z$

X	Y	Z
1..8	1..8	1..8
1..8	2..8	
	2..3	5..8
1..4	2..3	

Search + Propagate



Subject to min(Y-Z)

Strategy

1. Create problem variables, and
2. define domains.
2. Set up constraints
3. Propagate! Ensure consistency
4. Search + propagate (possibly subject to objective function).

Constraint Encoding

Let $s \in \{H,P\}^*$.

Problem variables+domains

For each $1 \le i \le |s|$:

- X_i in $1..2|s|$
- Y_i in $1..2 \cdot |s|$
- Z_i in $1..2 \cdot |s|$

For each $1 \le i < j \le |s|$:

- $\Delta X_{i,j}$ in $0..(j-i)$
- $\Delta Y_{i,j}$ in $0..(j-i)$
- $\Delta Z_{i,j}$ in $0..(j-i)$

For each $1 \le i+1 < j \le |s|$

- $\text{Contact}_{i,j}$ in $0..1$

Self-avoidance

For each $1 \le i < j \le |s|$, $(x_i, y_i, z_i) \neq (x_j, y_j, z_j)$

Or alternatively:

- $\Delta X_{i,j} = abs(x_i - x_j)$
- $\Delta Y_{i,j} = abs(y_i - y_j)$
- $\Delta Y_{i,j} = abs(z_i - z_j)$
- $\Delta X_{i,j} + \Delta Y_{i,j} + \Delta Z_{i,j} > 0$

Unit distance

For each $1 \le i < |s|$:

$$\|(x_i, y_i, z_i) - (x_{i+1}, y_{i+1}, z_{i+1})\| = 1$$

Or alternatively:

- $\Delta X_{i,i+1} = abs(x_i - x_{i+1})$
- $\Delta Y_{i,i+1} = abs(y_i - x_{i+1})$
- $\Delta Z_{i,i+1} = abs(z_i - z_{i+1})$
- $\Delta X_{i,i+1} + \Delta Y_{i,i+1} + \Delta Z_{i,i+1} = 1$

Contacts

For each $1 \le i$ and $j \le |s|$ where $i+1 < j$:

- $Contact_{i,j} \leftrightarrow (\Delta X_{i,j} + \Delta Y_{i,j} + \Delta Z_{i,j} = 1)$

Finally:

- $HHcontacts = \sum_{\substack{i+1<j \\ s(i)=s(j)=H}} Contact_{i,j}$
- $Energy = -HHcontacts$

Solve all constraints subject to

- $min(Energy)$

Optimizations (see paper)

- Redundant constraints (for more propagation)
- Bounds on energy. For example
  - An internal H-monomer can have at most 2 contacts in the 2D-case
  - An extremal H-monomer can have at most 3 contacts.
  - An even monomer can only have contact with odd monomers
- Try to enumerate low energy conformations early (more pruning during search in combination with bounds.)

Exercise

Draw the native conformation of the "protein" HHPPHHHHPPHPHH in the 2-dimensional HP-model. What is the energy of the conformation?

Hint: It should be possible to solve the problem using a rather naive algorithm.