CLP-BASED PROTEIN FRAGMENT ASSEMBLY

A. Dal Palù[1] A. Dovier[2] F. Fogolari[3] E. Pontelli[4]

1. Dept. Math, Univ. Parma

2. Dept. Math and Computer Science, Univ. Udine

3. Dept. Biomedical Sciences and Technologies, Univ. Udine

4. Dept. Computer Science, New Mexico State Univ.

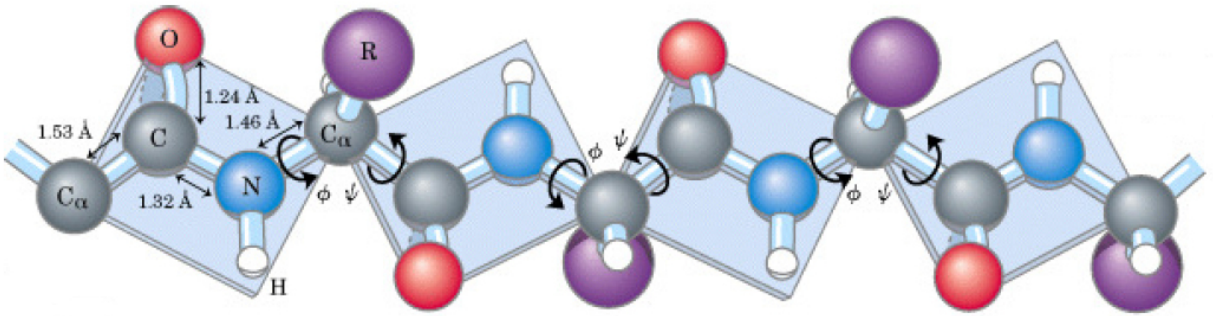
ICLP 2010

SUMMARY

PROTEINS

AMINO ACIDS

- Proteins are molecules made of amino acids
- Amino acids are small molecules that form a flexible chain

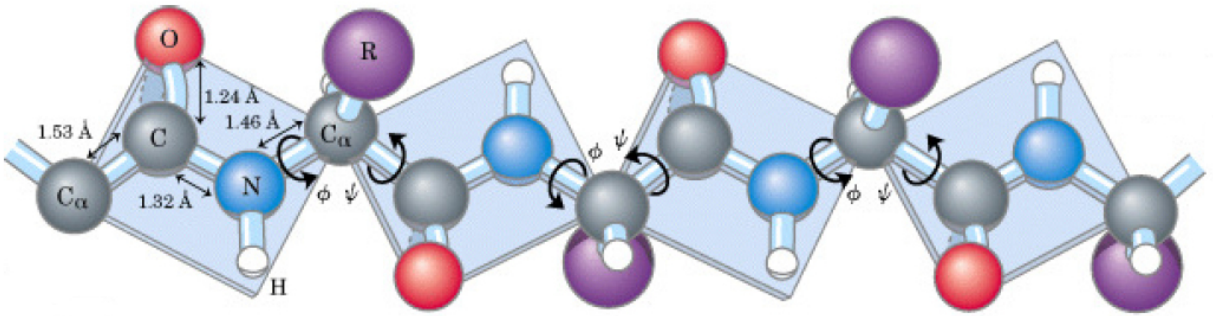


- There are several degrees of freedom (black arrows)
- The purple dots represent the *side chains*, that is a sub molecule that depends on the amino acid type
- Each amino acid is given a letter code (typ., 20 types)

PROTEINS

AMINO ACIDS

- Proteins are molecules made of amino acids
- Amino acids are small molecules that form a flexible chain

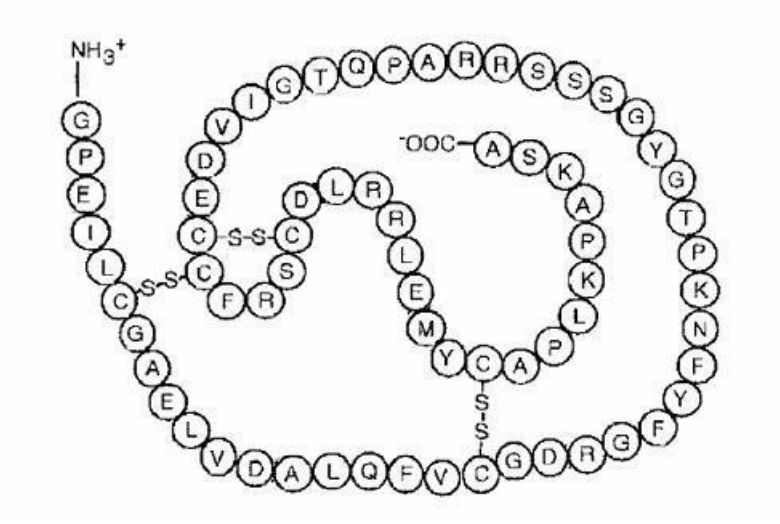


- There are several degrees of freedom (black arrows)
- The purple dots represent the *side chains*, that is a sub molecule that depends on the amino acid type
- Each amino acid is given a letter code (typ., 20 types)

PROTEINS

PRIMARY SEQUENCE

Aminoacid sequence:
GPEILCGAELVDALQFVCGDRGFYFNKPTGYGSSS
RRAPQTGIVDECCFRSCDLRRLEMYCAPLKPAKSA

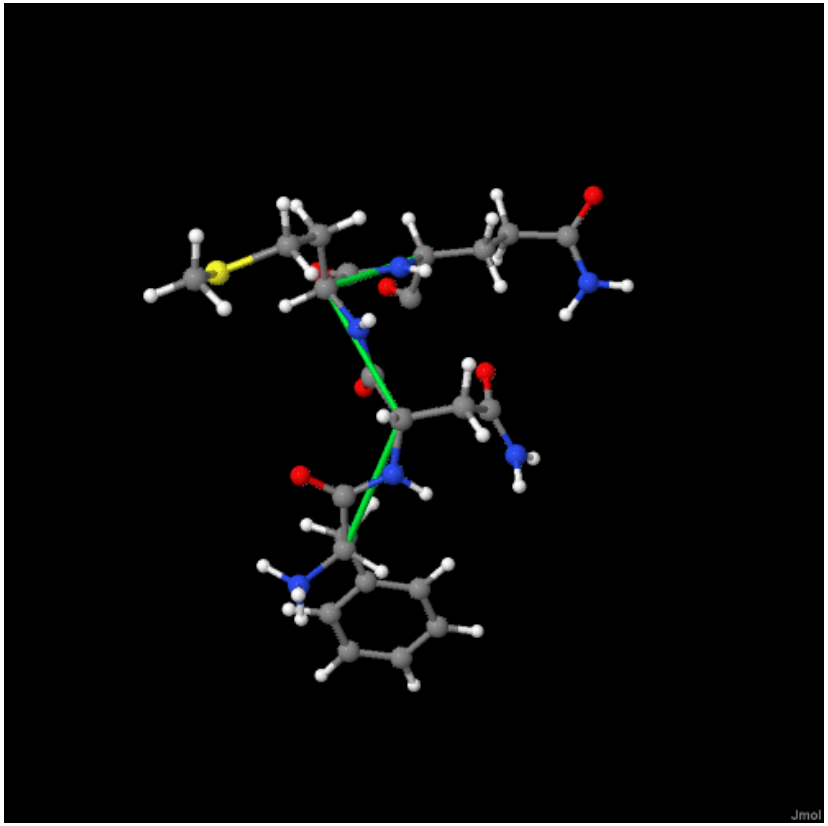


Primary sequence...

...embedded in the 3D space

PROTEINS

SIMPLIFIED ATOMIC STRUCTURE

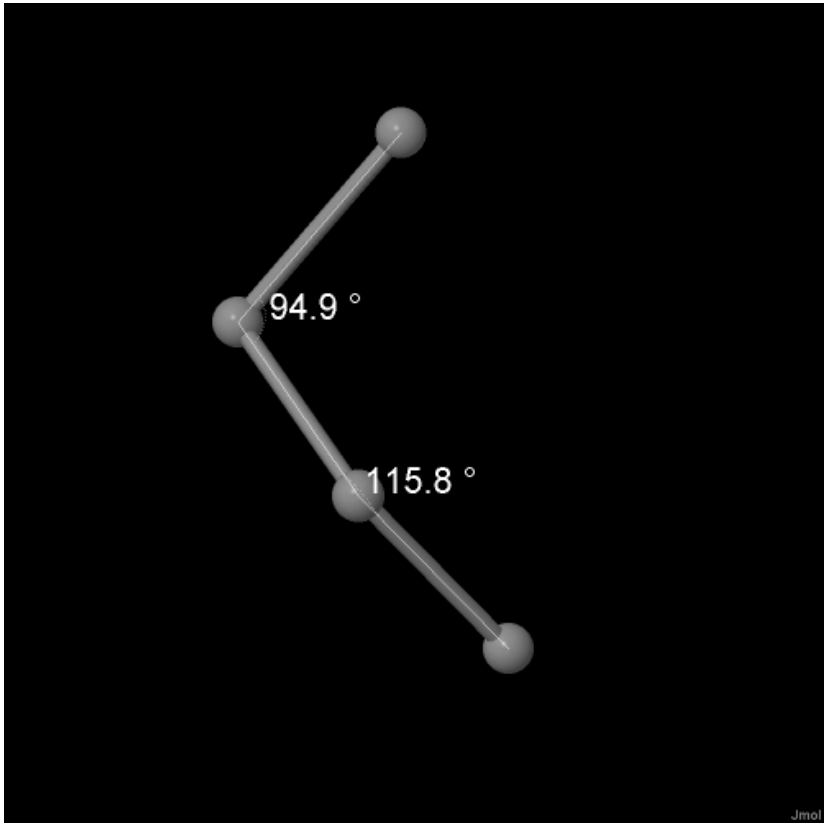


- The *backbone* (green) links consecutive amino acids
- Each amino acid is represented by its carbon alpha (C_α)

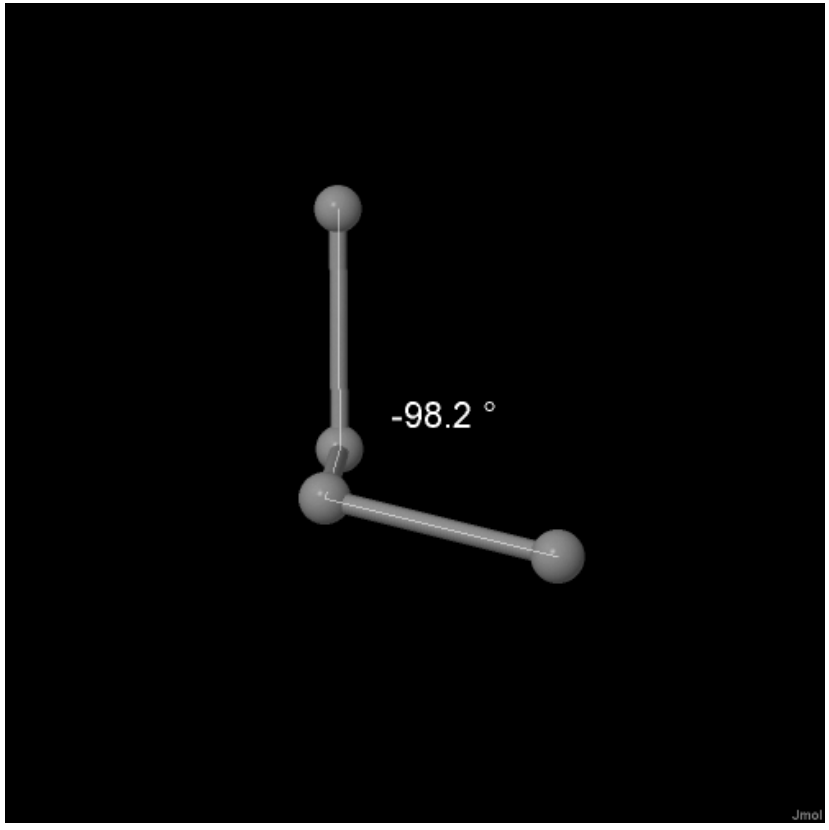
PROTEINS

BACKBONE

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



There are two bend angles



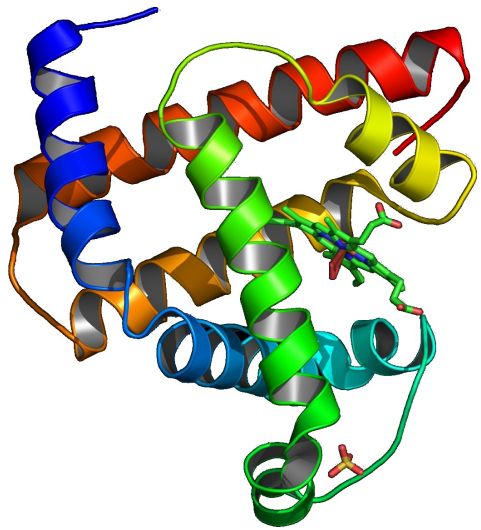
And one torsional angle

THE STRUCTURE PREDICTION PROBLEM

- Given the amino acid sequence
- For each amino acid, output its position in the space

A	L	F	W	K	L	R	R	...

? ⇓ ?



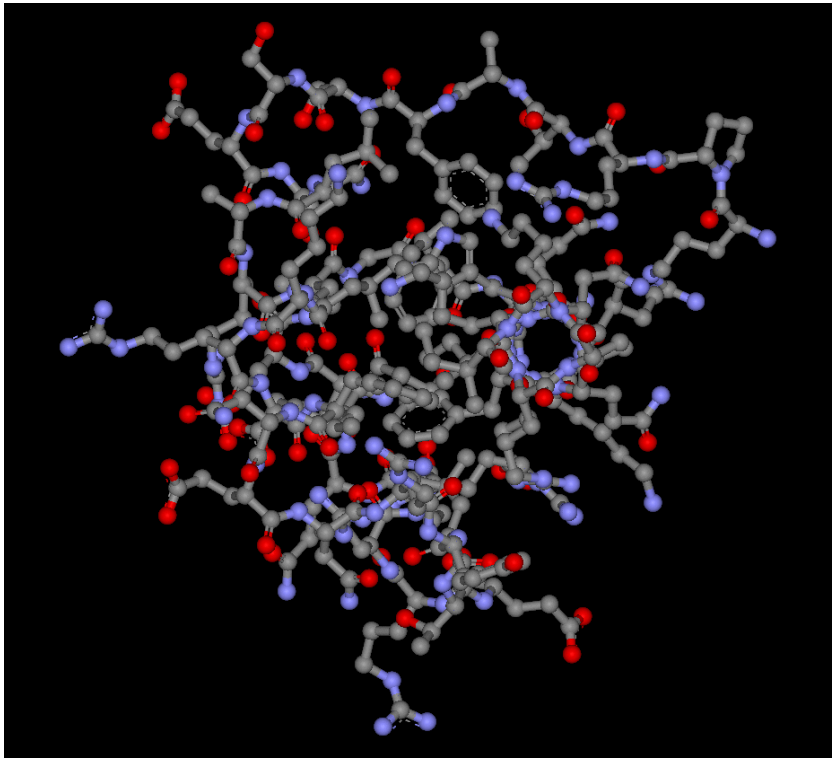
THE STRUCTURE PREDICTION PROBLEM

... and this is the hard part:

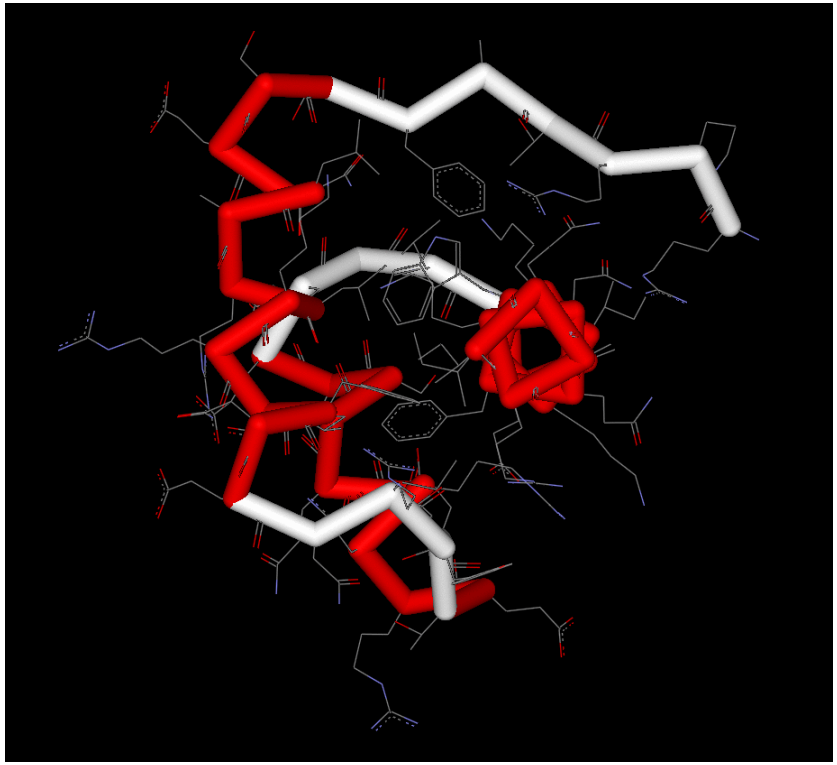
- In nature a protein has a unique/stable 3D conformation
- A cost function (that mimics physics laws) can be used to score each conformation
- Searching for the optimal score produces the best candidate (NP-complete)

PROTEINS

EXAMPLE



A full atom view



The backbone

IDEA: FRAGMENT ASSEMBLY

- Too many random 3D conformations: focus on biologically meaningful ones
- Generation of conformations constrained by *local homology*
- Similar sub-sequences $\Leftrightarrow$ similar shapes
- The shorter the sequence, the more typical behavior (set of most frequent conformations)
- Local shape (e.g., helices) depends more on local sequence
- Idea: enumeration of local shapes (fragment assembly)
- Combine local (statistical) preferences and build a coherent global conformation
- Assumption: nature prefers some local shapes $\Longrightarrow$ we should do it as well

IDEA: FRAGMENT ASSEMBLY

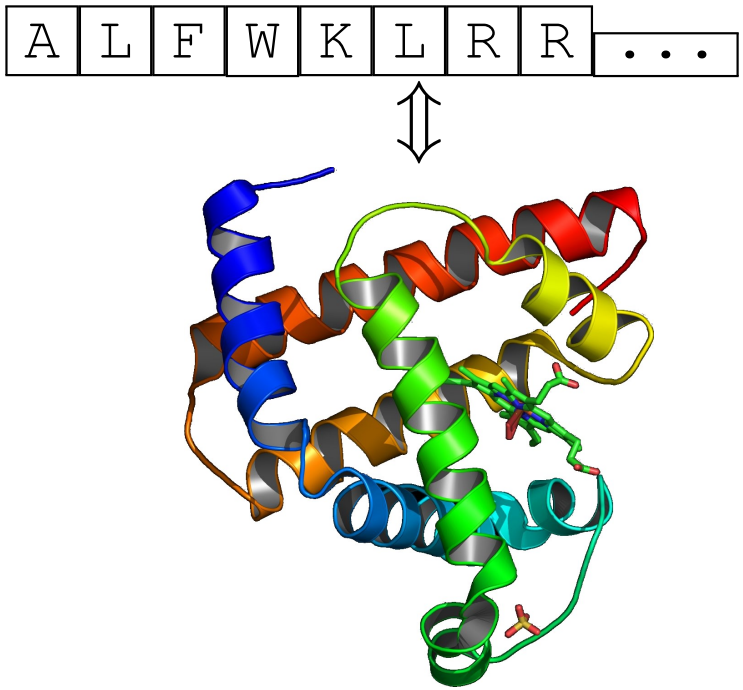
- Too many random 3D conformations: focus on biologically meaningful ones
- Generation of conformations constrained by *local homology*
- Similar sub-sequences $\Leftrightarrow$ similar shapes
- The shorter the sequence, the more typical behavior (set of most frequent conformations)
- Local shape (e.g., helices) depends more on local sequence
- Idea: enumeration of local shapes (fragment assembly)
- Combine local (statistical) preferences and build a coherent global conformation
- Assumption: nature prefers some local shapes $\Longrightarrow$ we should do it as well

IDEA: FRAGMENT ASSEMBLY

- Too many random 3D conformations: focus on biologically meaningful ones
- Generation of conformations constrained by *local homology*
- Similar sub-sequences $\Leftrightarrow$ similar shapes
- The shorter the sequence, the more typical behavior (set of most frequent conformations)
- Local shape (e.g., helices) depends more on local sequence
- Idea: enumeration of local shapes (fragment assembly)
- Combine local (statistical) preferences and build a coherent global conformation
- Assumption: nature prefers some local shapes $\Longrightarrow$ we should do it as well

PREPROCESSING

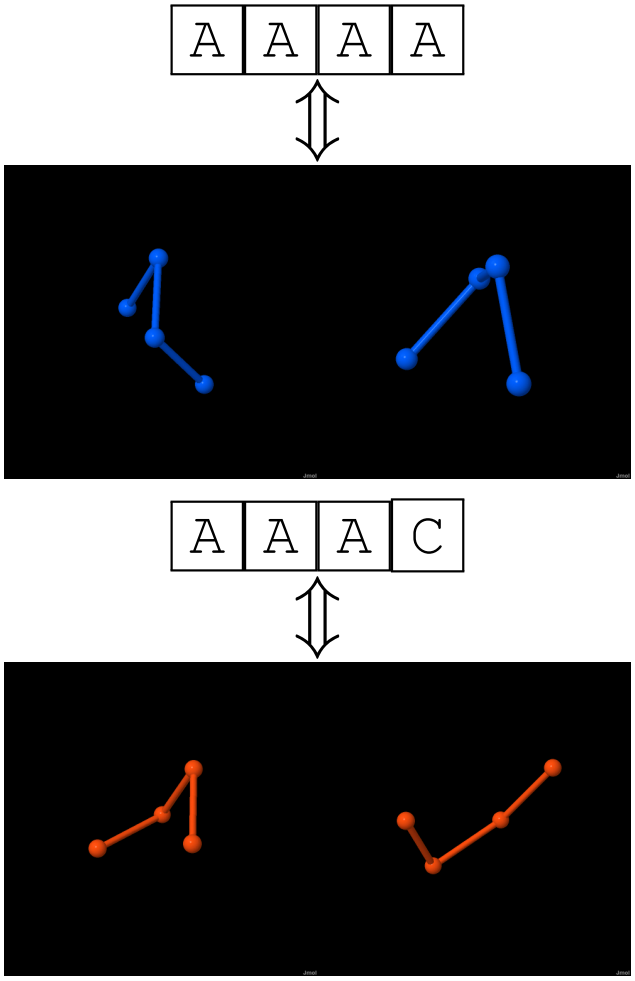
The Protein Data Bank contains $\geq 60K$ protein sequences with their observed 3D structures (X-ray/NMR)
Reaching 600K protein sequences (Swissprot)



PREPROCESSING

PDB: EXTRACT INFORMATION

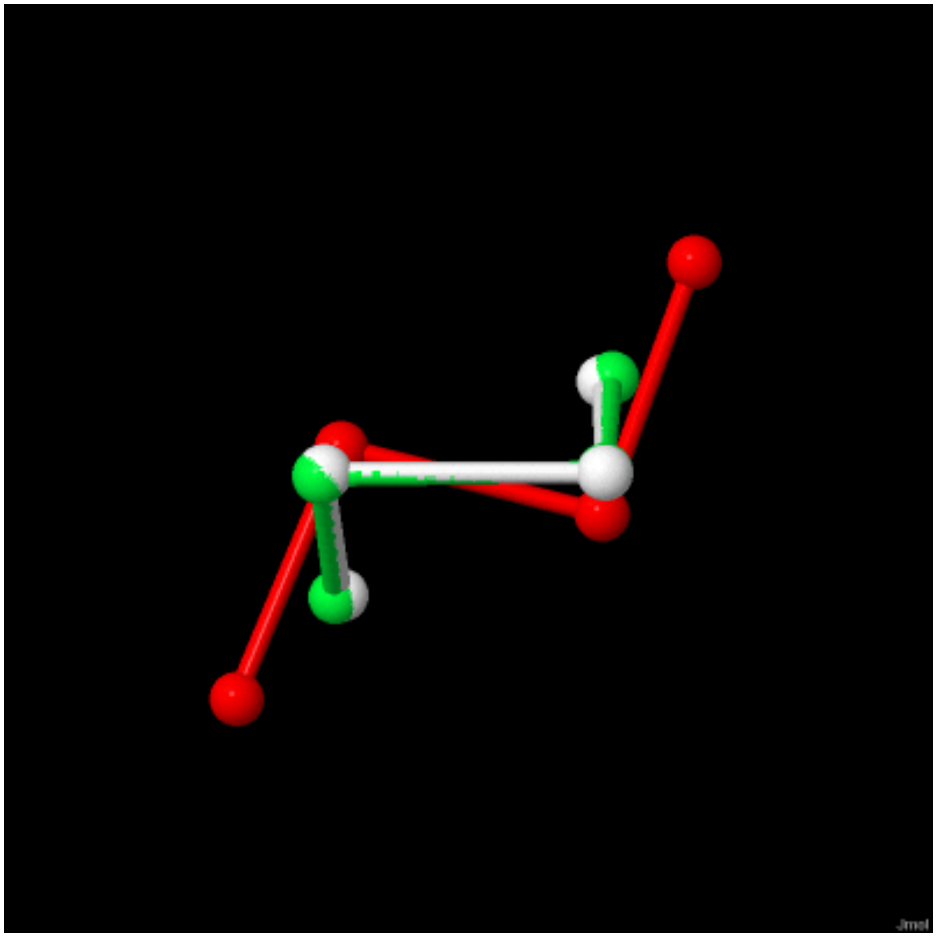
We get fragments composed of 4 consecutive amino acids and collect the corresponding shapes (indexed by sequence)



PREPROCESSING

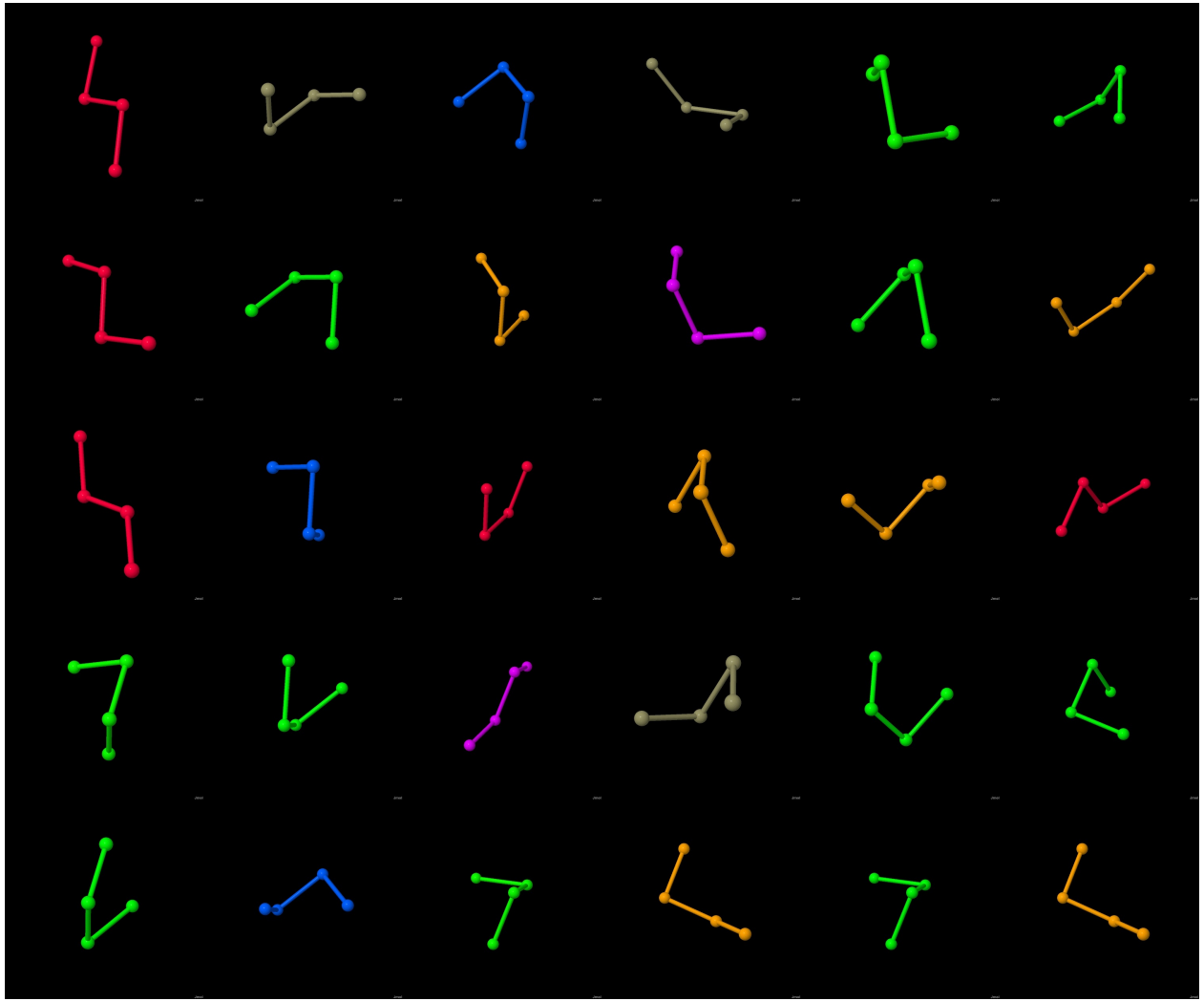
CLUSTERING (SAME 4-PLE, DIFFERENT SHAPES)

Clustering according to their similarity (RMSD $\leq$ threshold)
White and green form a single cluster



PREPROCESSING

CLUSTERED CONFORMATIONS FOR AAAA



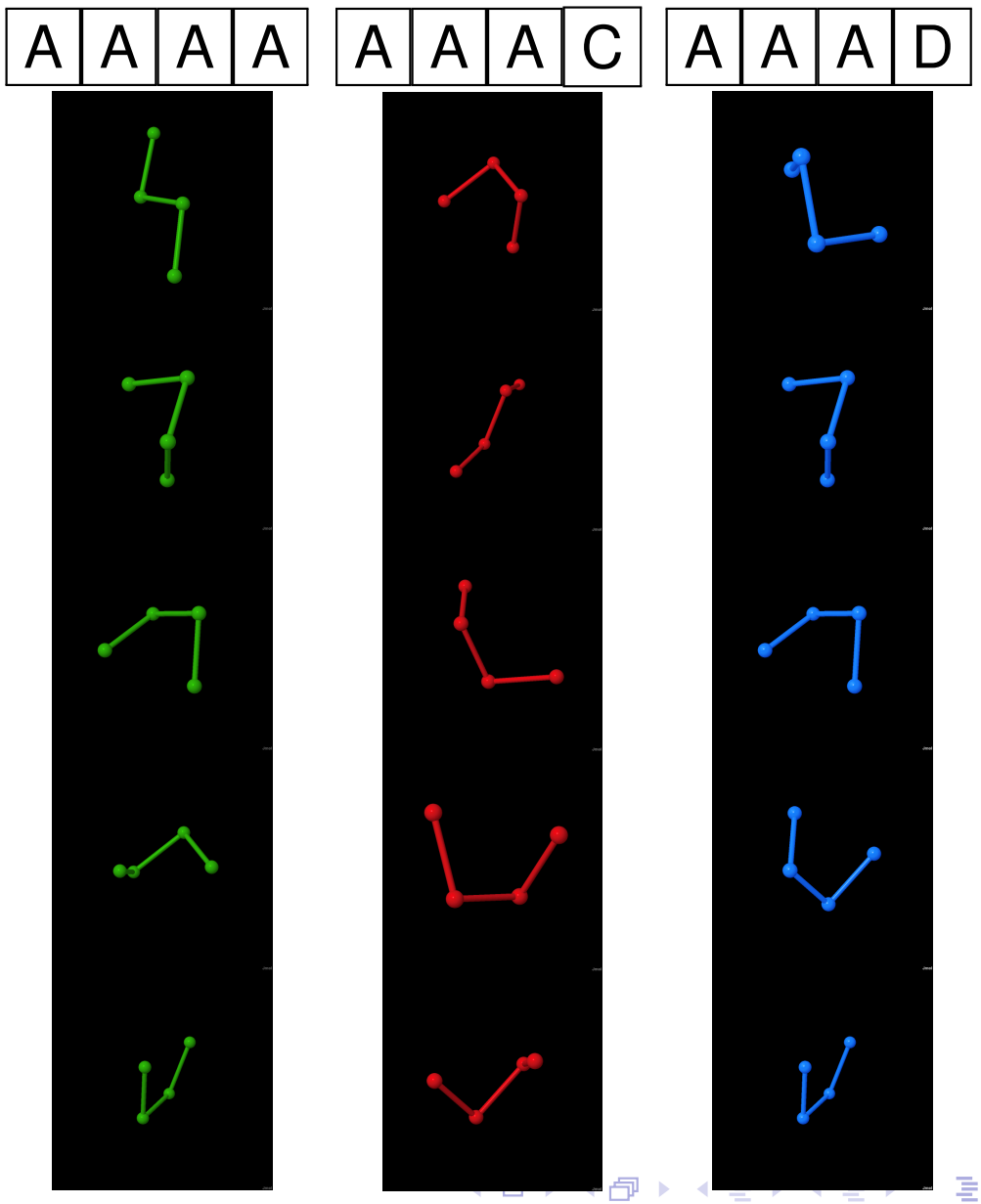
Each color has a representative and frequency count

PREPROCESSING

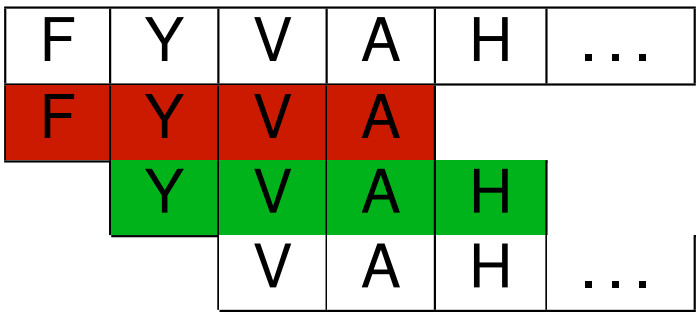
LIBRARY OF FRAGMENTS

For each 4 aa sequence, store the clustered representatives (RMSD $\leq$ .5Å) as a Prolog fact of the form

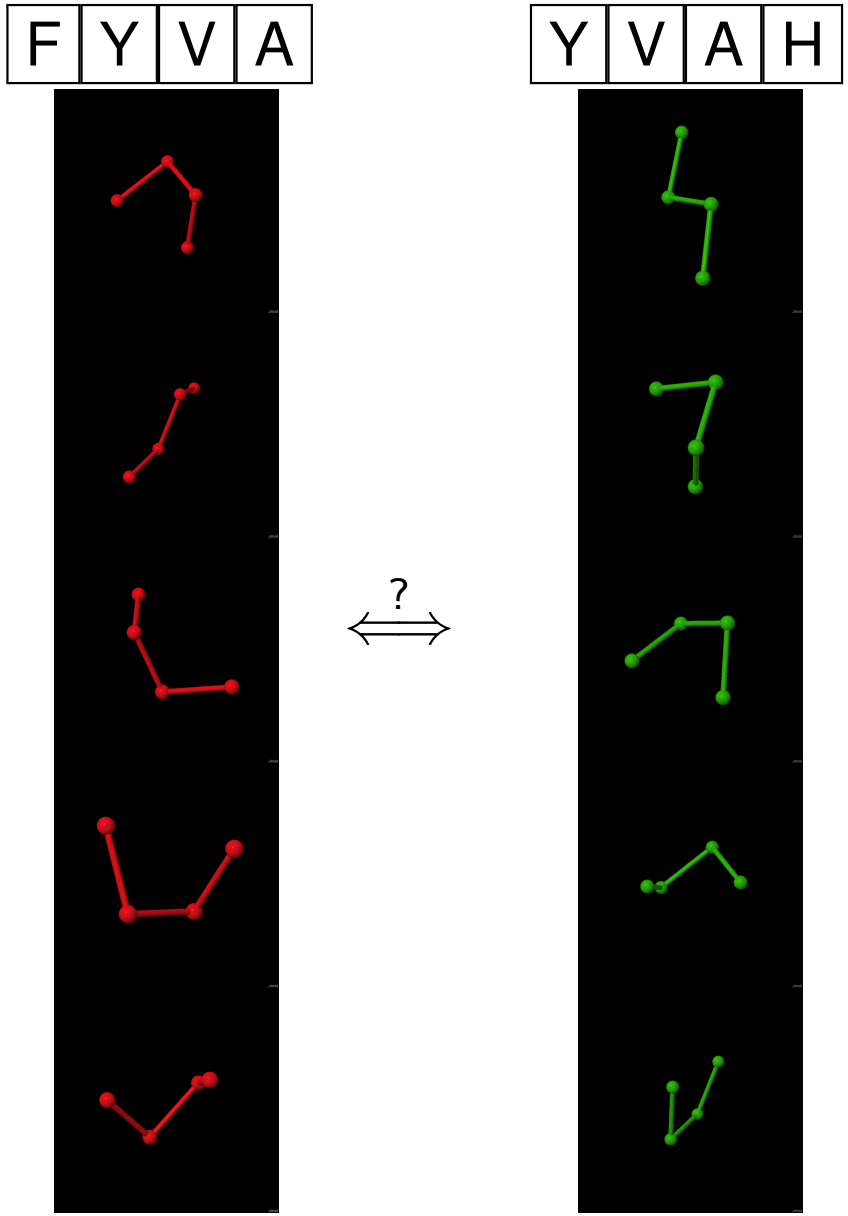
```
tupla([A,A,A,A],
[0.0,0.0,0.0,
 2.5,-2.8,0.3,
 1.9,-3.1,4.0,
 -1.9,-3.4,3.6],
 Freq, ID).
```



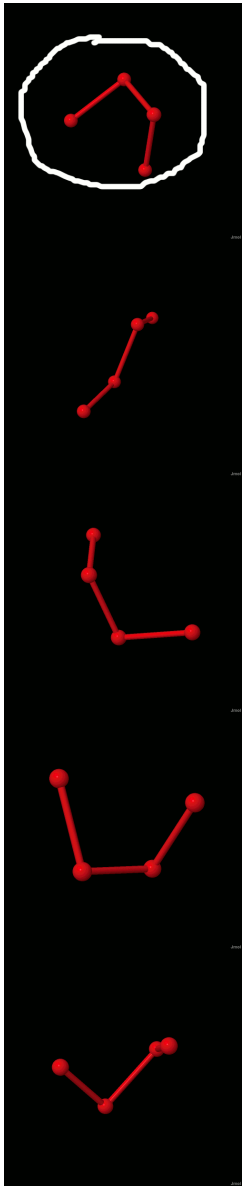
INDUCTIVE STEP: COMBINE THE BLOCKS

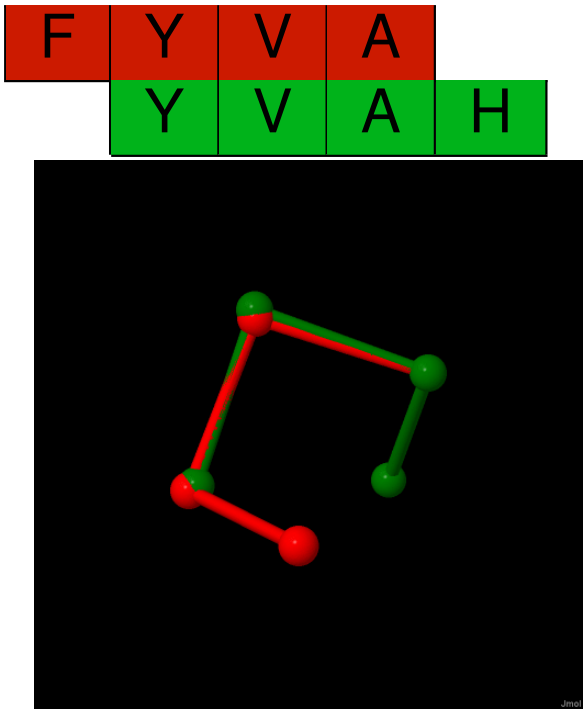


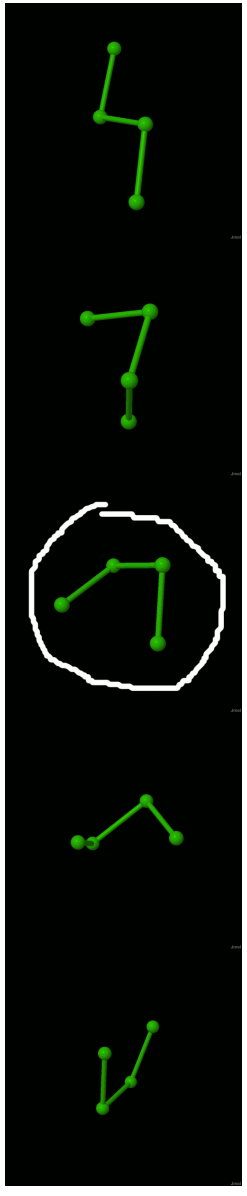
How to assemble fragments?



INDUCTIVE STEP: COMBINE THE BLOCKS

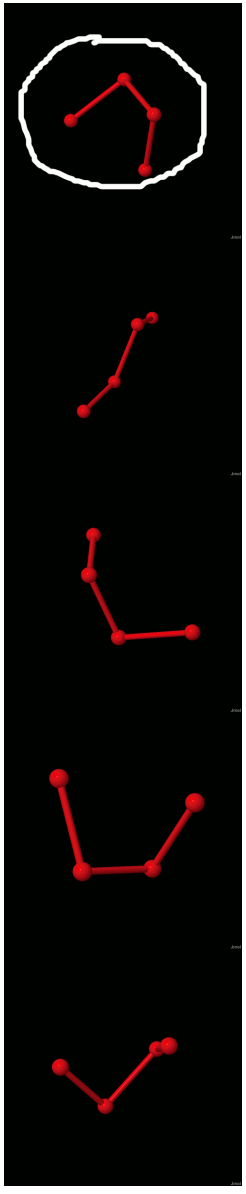


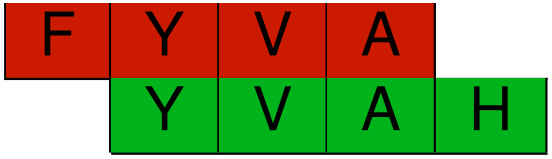




Two fragments are *compatible* only if the 3 common amino acids have a low RMSD (similar bend angle)

INDUCTIVE STEP: COMBINE THE BLOCKS

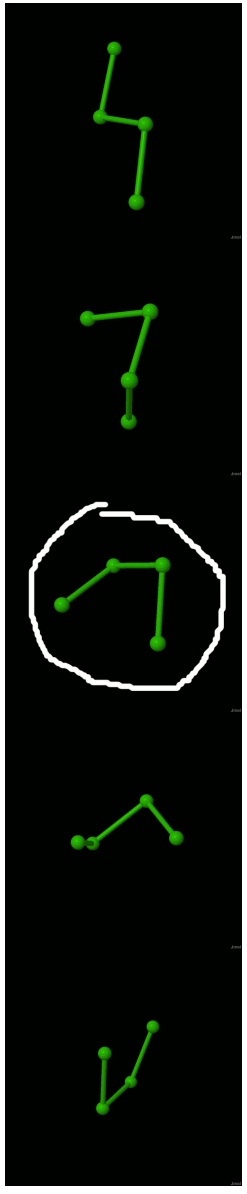




Each compatible pair of fragments is stored as

$$next(F_i, F_j, M)$$

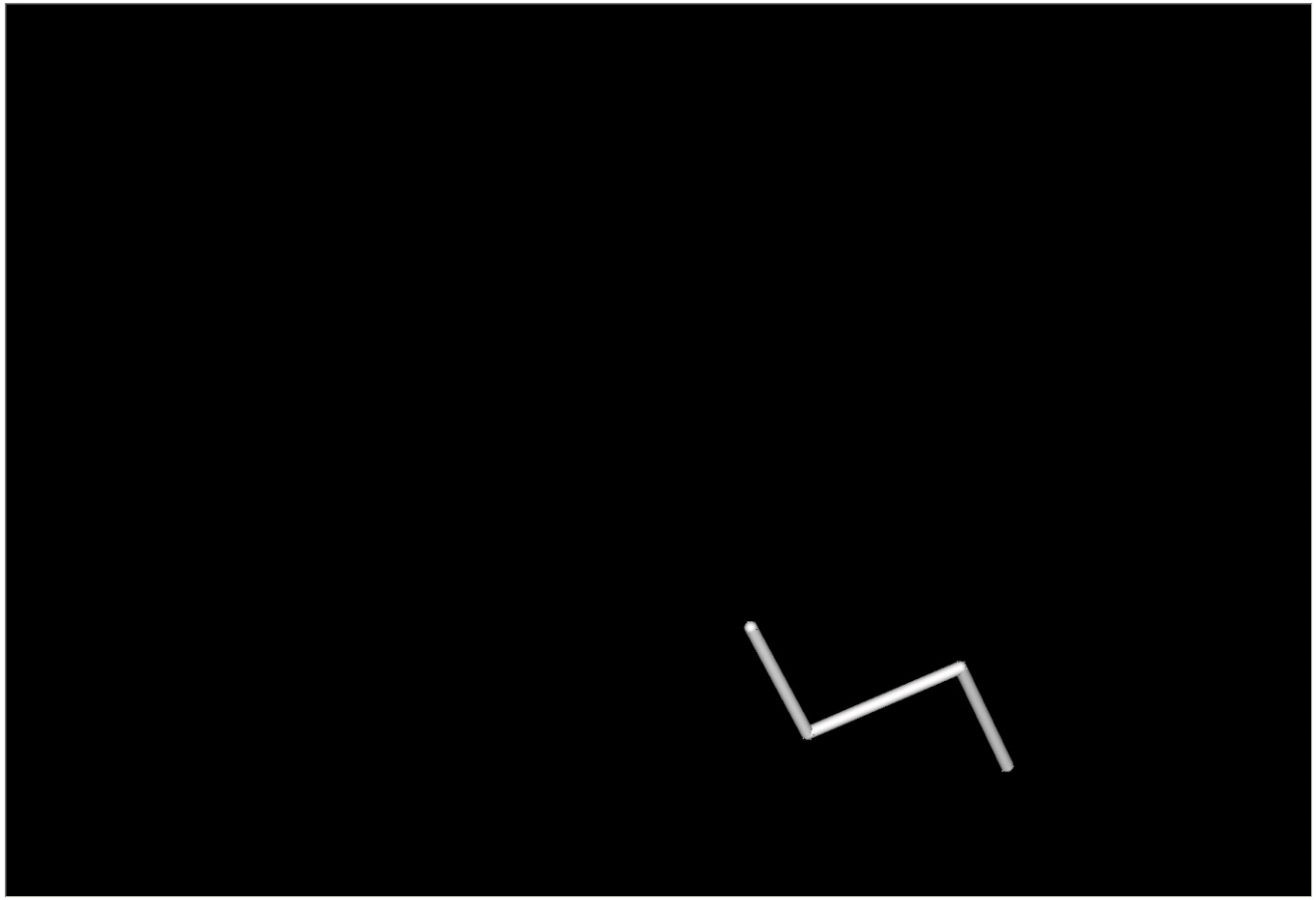
with optimal rotation matrix M (that rotates F_j in the reference of F_i)



INDUCTIVE STEP: COMBINE THE BLOCKS

THE ASSEMBLY

Given a target sequence, pick the first 4-aa fragment.
The protein is grown by attaching compatible fragments (*next*).



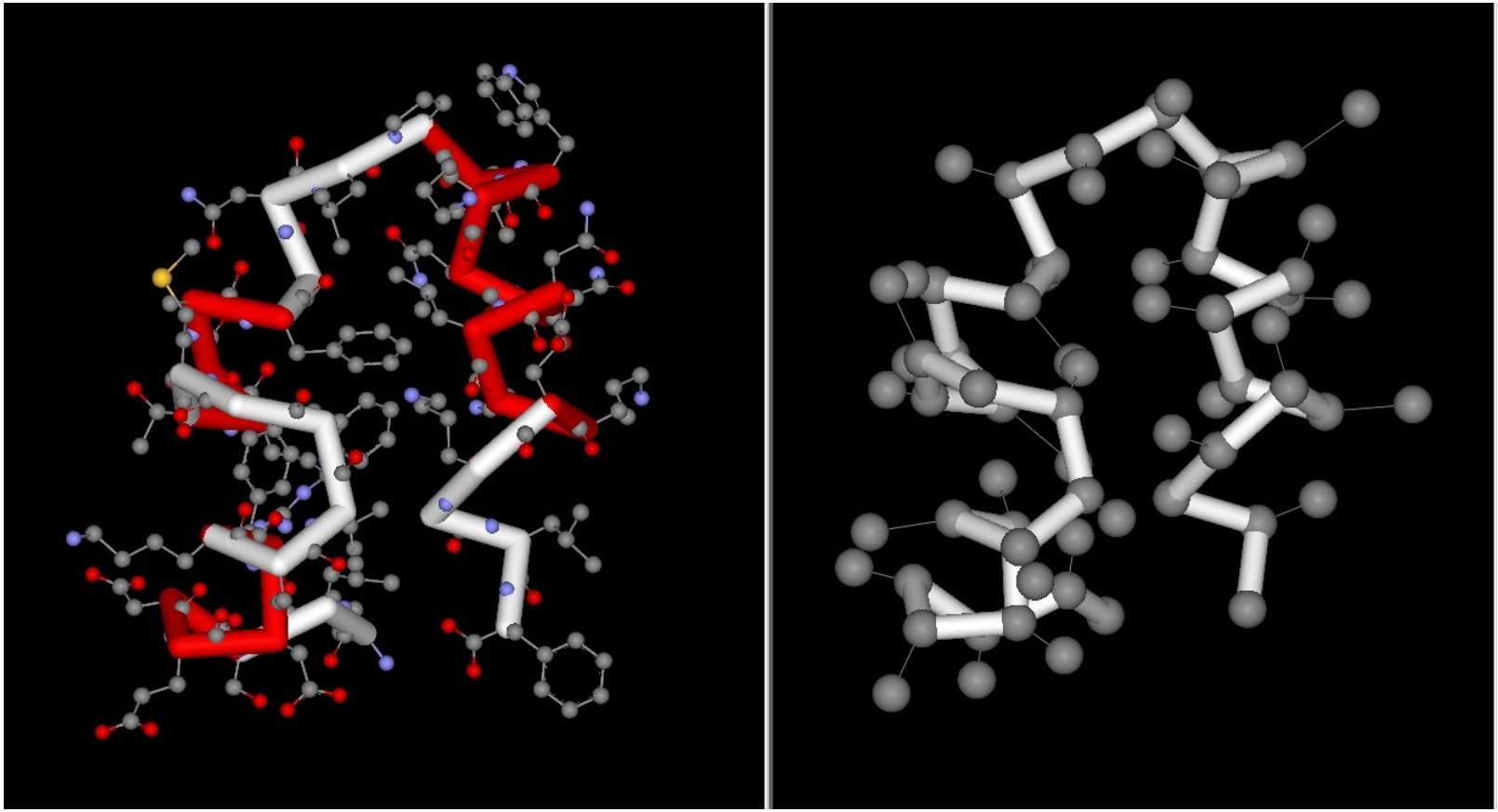
ENRICHING THE MODEL

THE CENTROIDS

- Given a $C\alpha$ 4-tuple in 3D, a small degree of freedom for the position of the side chain is allowed
- Different amino acids have different occupation
- A pure $C\alpha$-$C\alpha$ model does not keep into account these differencies
- We consider the positions of the centroids of the side chains.
- Roughly, a centroid is the expected center of mass of the side chain

ENRICHING THE MODEL

THE CENTROIDS



CLP(FD) MODELING

FD VARIABLES

- Given amino acid sequence: $[a_1, a_2, \ldots, a_n]$,
- Amino acid positions: $X_1^\alpha, Y_1^\alpha, Z_1^\alpha, \ldots, X_n^\alpha, Y_n^\alpha, Z_n^\alpha, X_2^c, Y_2^c, Z_2^c, \ldots, X_{n-1}^c, Y_{n-1}^c, Z_{n-1}^c$
- Domains: discretized (0.01Å) 3D positions.
- Fragments: $F_1, F_2, \ldots, F_{n-3}$.
- Domains: fragments ID from the library for $[a_i, a_{i+1}, a_{i+2}, a_{i+3}]$.
- Variables for rotation matrices: $R_1, R_2, \ldots, R_{n-3}$.

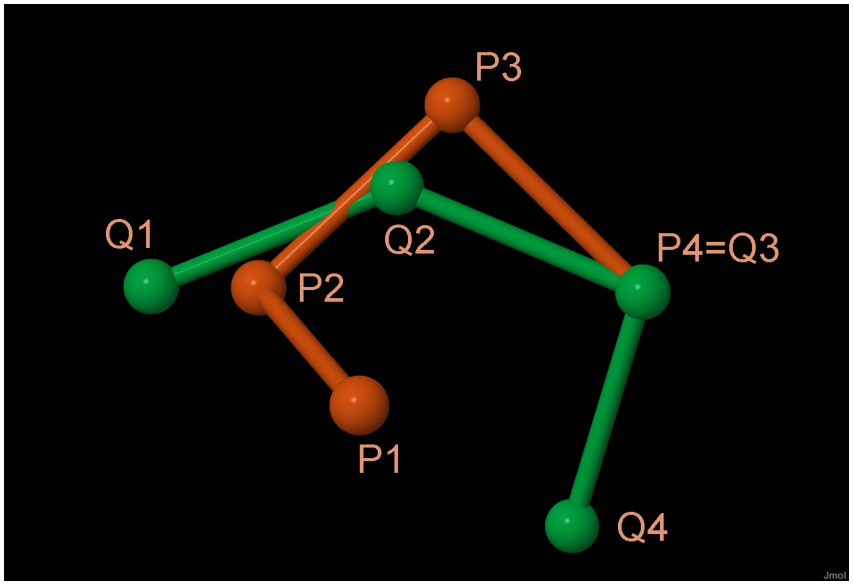
CLP(FD) MODELING

FD CONSTRAINTS

- F_i and F_{i+1} can take only compatible assignments (described by $next(F_i, F_{i+1}, M_{i,(i+1)})$)
- the constraint is posted using `table` builtin
- All fragments are stored in the same reference, thus need to rotate reference, while building the fragments:
- $R_{i+1} = M_{i,(i+1)} \cdot R_i$ implemented as a constraint

CLP(FD) MODELING

FD CONSTRAINTS: LINK POSITIONS TO FRAGMENTS



- $Q_4 = R_{i+1} \cdot F_{i+1} + \mathit{shift}$ (Ps already placed $= F_i$)
- correct shift overlaps P_4 and Q_3, so that distance between $(X^{\alpha}_{i+3}, Y^{\alpha}_{i+3}, Z^{\alpha}_{i+3})$ and $(X^{\alpha}_{i+4}, Y^{\alpha}_{i+4}, Z^{\alpha})$ is $\sim$3.81Å.

CLP(FD) MODELING

DISTANCE CONSTRAINTS

- For each i, j, introduce a distance var $D_{i,j}$
- $D_{i,j} = (X_i - X_j)^2 + (Y_i - Y_j)^2 + (Z_i - Z_j)^2$
- $D_{i,j} \geq$ *min_dist* (all_distant) — (*)
- $D_{i,j} \leq$ `diameter`^{2} (compact_factor)
- If ssbond(i,j), $D_{i,j} \leq$ 6Å.
- $D_{i,j} \leq 3.81 * |i - j|$ (max stretch)
- Triangular inequality (over $D_{i,j}$, $D_{i,k}$, $D_{j,k}$)

CLP(FD) MODELING

DISTANCE CONSTRAINTS

- For each i, j, introduce a distance var $D_{i,j}$
- $D_{i,j} = (X_i - X_j)^2 + (Y_i - Y_j)^2 + (Z_i - Z_j)^2$
- $D_{i,j} \geq$ *min_dist* (all_distant) — (*)
- $D_{i,j} \leq \texttt{diameter}^2$ (compact_factor)
- If ssbond(i,j), $D_{i,j} \leq 6$Å.
- $D_{i,j} \leq 3.81 * |i - j|$ (max stretch)
- Triangular inequality (over $D_{i,j}$, $D_{i,k}$, $D_{j,k}$)

CLP(FD) MODELING

COST FUNCTION

- Each conformation has four energy contributions:
- *Contact* centr–centr: each pair of centroid gives a contribution depending on distance and type.
- *Contact* $C\alpha$–$C\alpha$: each pair of centroid gives a contribution depending on distance.
- *Torsional*: each torsion is scored according to a statistical profile for the local sequence.
- *Torsional correlation*: each pair of consecutive torsion angles is scored with a correlation contribution.
- This cost function is new and developed with a researcher in molecular biology

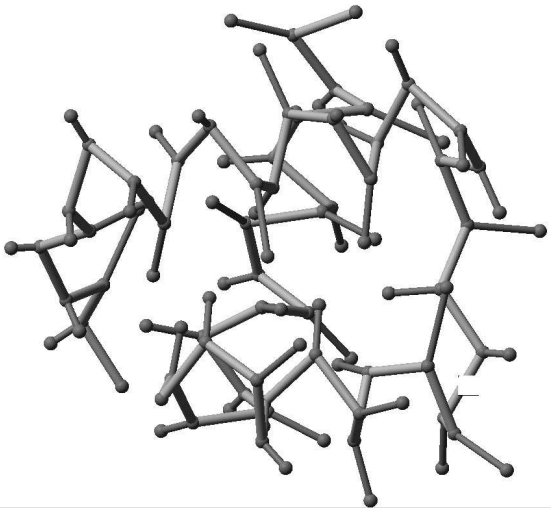
CLP(FD) MODELING

COST FUNCTION

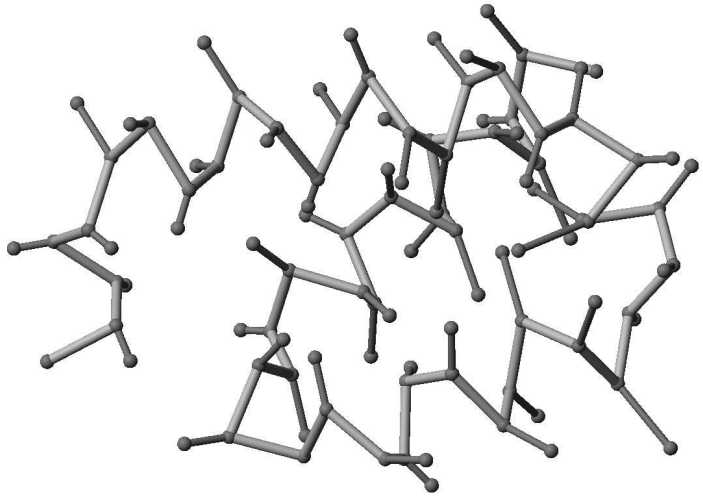
- Each conformation has four energy contributions:
- *Contact* centr–centr: each pair of centroid gives a contribution depending on distance and type.
- *Contact* $C\alpha$–$C\alpha$: each pair of centroid gives a contribution depending on distance.
- *Torsional*: each torsion is scored according to a statistical profile for the local sequence.
- *Torsional correlation*: each pair of consecutive torsion angles is scored with a correlation contribution.
- This cost function is new and developed with a researcher in molecular biology

RESULTS

1ENH



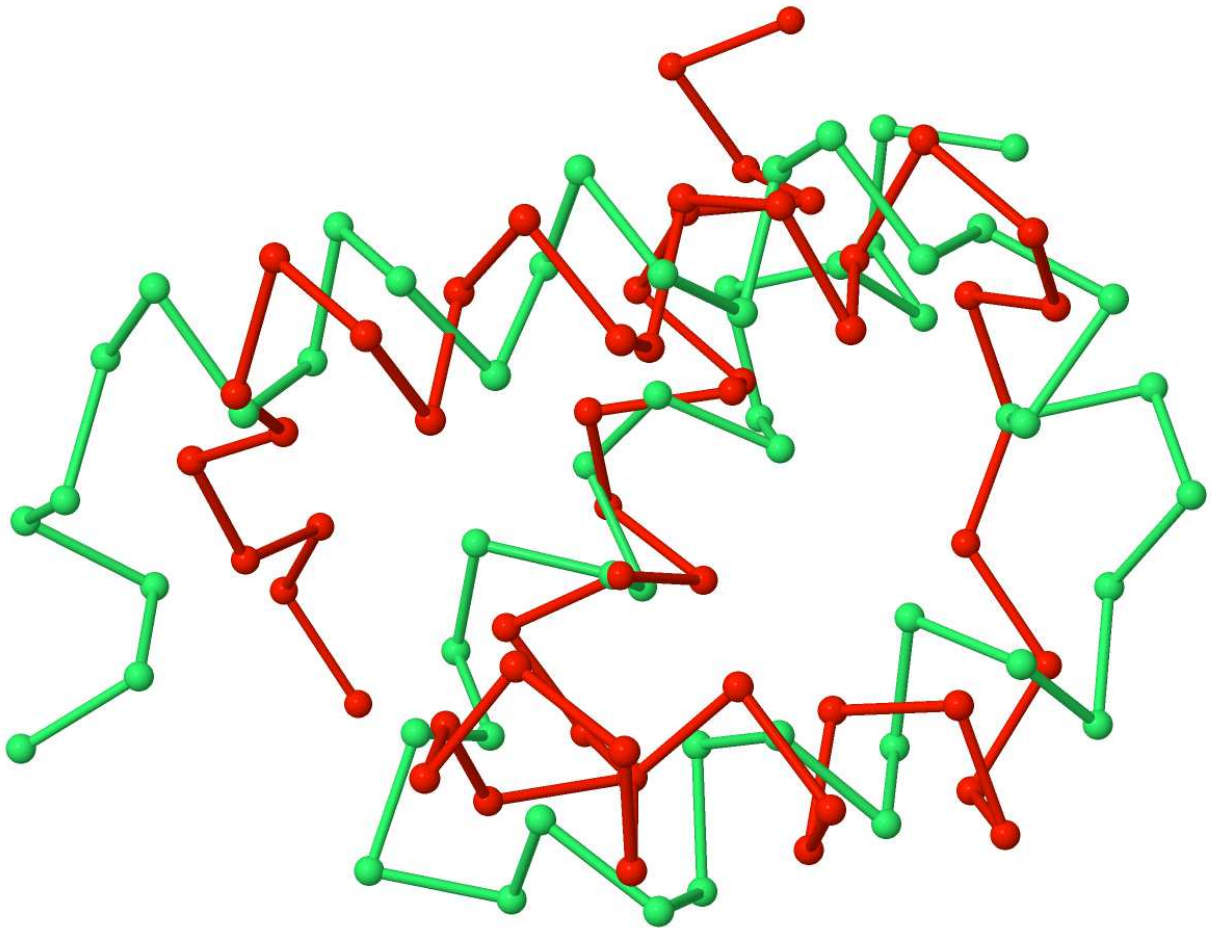
Original



Computed

RESULTS

1ENH



LARGE NEIGHBORING SEARCH

- Logic programming implementation of Large Neighboring Search
- LNS is a form of local search
- The search for next solutions is performed by exploring a *large* neighborhood
- A large number of variables (subject to constraints) is allowed to change.
- Use Prolog's backtracking capability to modify CLP variables (while constraints are not changed)

LARGE NEIGHBORING SEARCH

- Logic programming implementation of Large Neighboring Search
- LNS is a form of local search
- The search for next solutions is performed by exploring a *large* neighborhood
- A large number of variables (subject to constraints) is allowed to change.
- Use Prolog's backtracking capability to modify CLP variables (while constraints are not changed)

LARGE NEIGHBORING SEARCH

THE ALGORITHM

1. Generate an initial solution (standard CLP labeling).
2. Randomly select a subset of the variables and assign the previous values to the other variables (same fragments).
3. Standard labeling: look for an assignment that improves the cost function.
4. Go back to step 2.

Search stops with a timeout.
Allow occasional worse solutions (escape local optima)

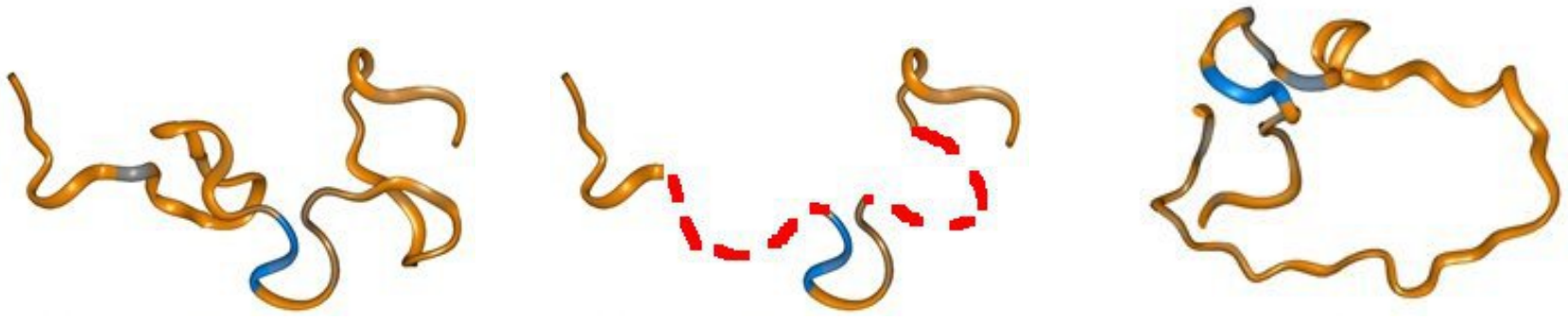
LARGE NEIGHBORING SEARCH

THE ALGORITHM

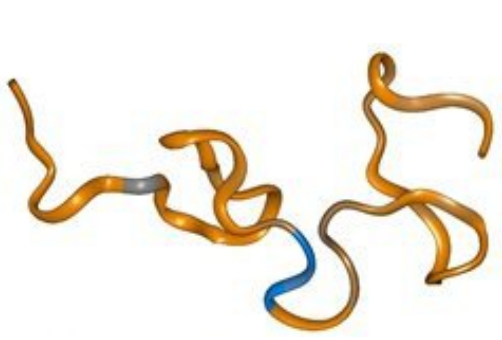
1. Generate an initial solution (standard CLP labeling).
2. Randomly select a subset of the variables and assign the previous values to the other variables (same fragments).
3. Standard labeling: look for an assignment that improves the cost function.
4. Go back to step 2.

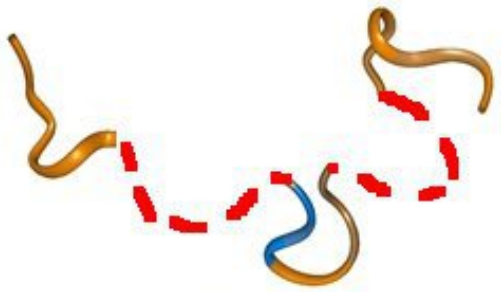
Search stops with a timeout.
Allow occasional worse solutions (escape local optima)

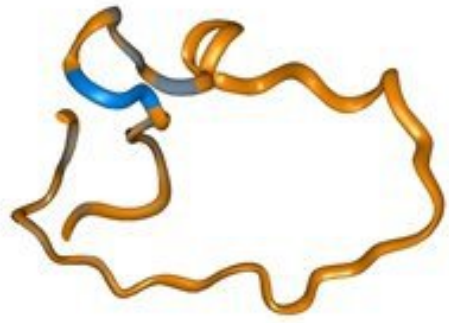
LARGE NEIGHBORING SEARCH

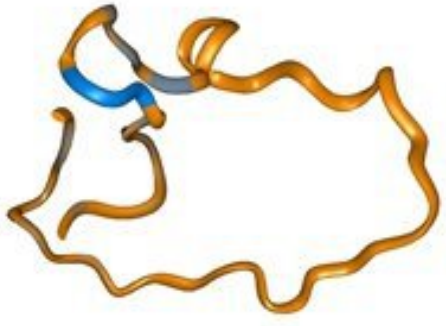


LARGE NEIGHBORING SEARCH

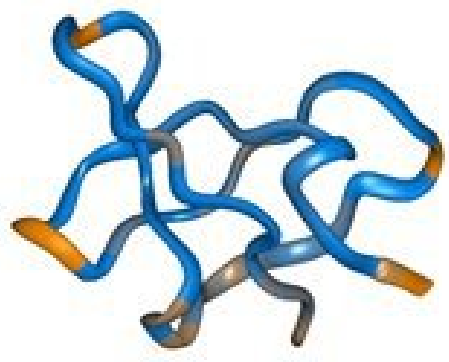








...



RESULTS

LARGE NEIGHBORING SEARCH (LNS) IN PROLOG

PID	N	Enumerate 2 days		LNS 2 hours	
		RMSD	Energy	RMSD	Energy
1ZDD	34	4.12	-231469	3.81	-226387
1AIL	69	9.78	-711302	5.53	-665161
1VII	36	7.06	-263496	6.64	-252231
2IGD	60	16.35	-375906	10.91	-447513

CONCLUSIONS

+ Working tool Prolog + CLP(FD) for PSP using the idea of tuple assembling
+ LNS (in Prolog) sensibly helps the search
+ We have developed a statistical potential function
+ The code is ready to deal with longer local predictions (it already deals with secondary structures)
- Correlation Energy/RMSD must be improved
- Constraint propagation must be improved (inverse kinematics?)

CONCLUSIONS

+ Working tool Prolog + CLP(FD) for PSP using the idea of tuple assembling
+ LNS (in Prolog) sensibly helps the search
+ We have developed a statistical potential function
+ The code is ready to deal with longer local predictions (it already deals with secondary structures)
- Correlation Energy/RMSD must be improved
- Constraint propagation must be improved (inverse kinematics?)

WCB 2010

Workshop on
Constraint Based Methods for Bioinformatics

July 21st, 2010 Edinburgh
colocated to ICLP 2010

Organizers:

- Alessandro Dal Palù
- Agostino Dovier
- Sebastian Will