Candidate Fragments Prediction and their Assembly with a Greedy Algorithm and a Coarse-Grained Force Field to solve Protein Folding

JOBIM 2007

Julien Maupetit Frédéric Guyon Anne-Claude Camproux Philippe Derreumaux Pierre Tufféry

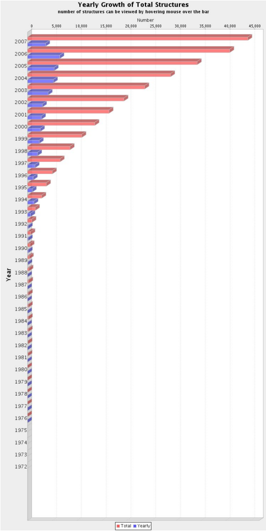
EBGM - Equipe Bioinformatique Génomique et Moléculaire
INSERM U726 - Université Denis Diderot Paris 7
FRANCE

2007/07/11

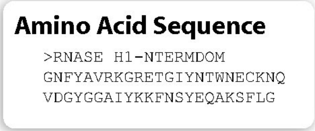
Introduction

- **Sequence** databases grow exponentially.
- ~ **20-25 % of orphan genes.**
- **Comparative modeling** approaches are very accurate, but not for orphan genes.

↪ ***ab initio* / *de novo* methods**



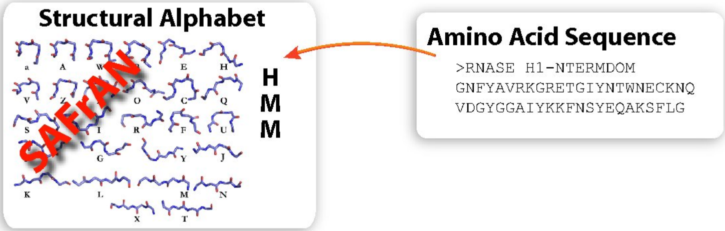
The HMM-SA method



Amino Acid Sequence

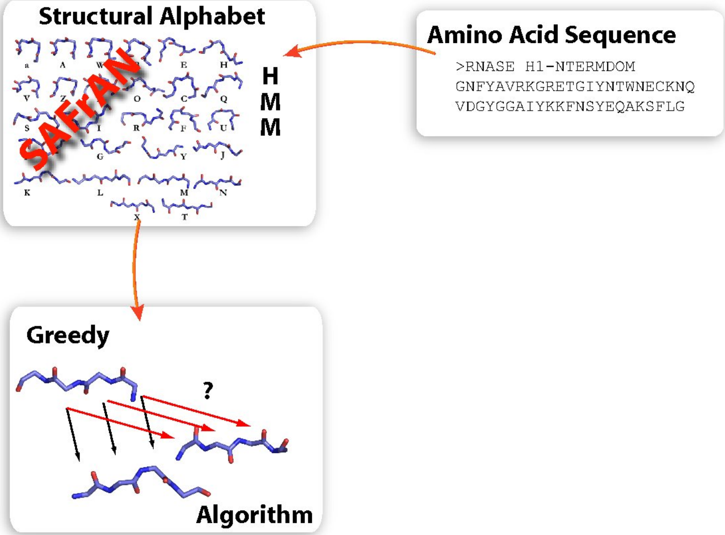
```
>RNASE H1-NTERMDOM
GNFYAVRKGRETGIYNTWNECKNQ
VDGYGGAIYKKFNSYEQAKSFLG
```

The HMM-SA method



```
>RNASE H1-NTERMDOM
GNFYAVRKGRETGIYNTWNECKNQ
VDGYGGAIYKKFNSYEQAKSFLG
```

The HMM-SA method



Structural Alphabet

H
M
M

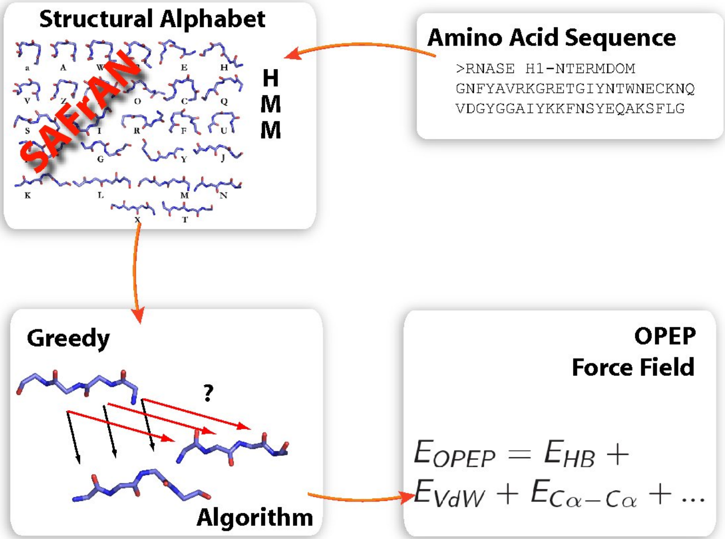
Amino Acid Sequence

>RNASE H1-NTERMDOM
GNFYAVRKGRETGIYNTWNECKNQ
VDGYGGAIYKKFNSYEQAKSFLG

Greedy

Algorithm

The HMM-SA method



Structural Alphabet

H M M

Amino Acid Sequence

```
>RNASE H1-NTERMDOM
GNFYAVRKGRETGIYNTWNECKNQ
VDGYGGAIYKKFNSYEQAKSFLG
```

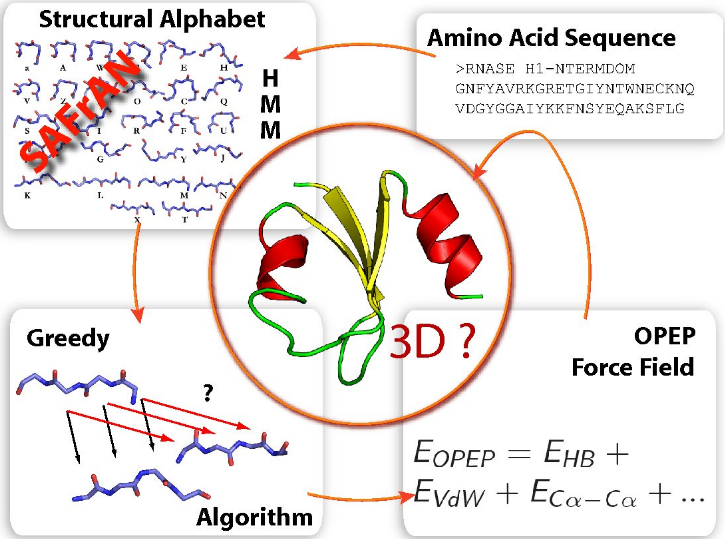
Greedy

Algorithm

OPEP Force Field

$$E_{OPEP} = E_{HB} + E_{VdW} + E_{C\alpha - C\alpha} + \ldots$$

The HMM-SA method



Structural Alphabet

Amino Acid Sequence

```
>RNASE H1-NTERMDOM
GNFYAVRKGRETGIYNTWNECKNQ
VDGYGGAIYKKFNSYEQAKSFLG
```

Greedy

Algorithm

OPEP

Force Field

$$E_{OPEP} = E_{HB} + E_{VdW} + E_{C\alpha-C\alpha} + \ldots$$

Outline

1. HMM based Structural Alphabet
   - HMM-SA27
   - Structural Alphabet (SA)

Outline

1. HMM based Structural Alphabet
   - HMM-SA27
   - Structural Alphabet (SA)
2. SAFrAN
   - SA prediction from amino acids sequence
   - SA Search
   - SAFrAN algorithm
   - SAFrAN example
   - Candidate fragments properties

Outline

1. HMM based Structural Alphabet
   - HMM-SA27
   - Structural Alphabet (SA)
2. SAFrAN
   - SA prediction from amino acids sequence
   - SA Search
   - SAFrAN algorithm
   - SAFrAN example
   - Candidate fragments properties
3. Greedy-OPEP
   - Greedy algorithm
   - OPEP: a coarse-grain force field

Outline

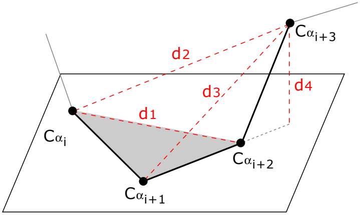
1. HMM based Structural Alphabet
   - HMM-SA27
   - Structural Alphabet (SA)
2. SAFrAN
   - SA prediction from amino acids sequence
   - SA Search
   - SAFrAN algorithm
   - SAFrAN example
   - Candidate fragments properties
3. Greedy-OPEP
   - Greedy algorithm
   - OPEP: a coarse-grain force field
4. Results - Discussion
   - CASP7 experiment
   - Improvements since CASP7

Outline

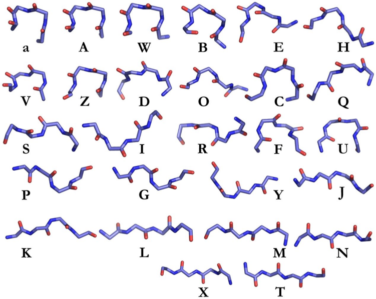
1. HMM based Structural Alphabet
   - HMM-SA27
   - Structural Alphabet (SA)
2. SAFrAN
   - SA prediction from amino acids sequence
   - SA Search
   - SAFrAN algorithm
   - SAFrAN example
   - Candidate fragments properties
3. Greedy-OPEP
   - Greedy algorithm
   - OPEP: a coarse-grain force field
4. Results - Discussion
   - CASP7 experiment
   - Improvements since CASP7

HMM-SA27

HMM-SA descriptors:



HMM-SA 27 states:



HMM-SA Properties

- 1 letter is **4 residues length** protein fragment
- Overlap on 3 residues
- HMM descriptors: $d_1 d_2 d_3 d_4$
- Learnt from 1429 PDB structures
- **27 HMM states** (155 prototypes)

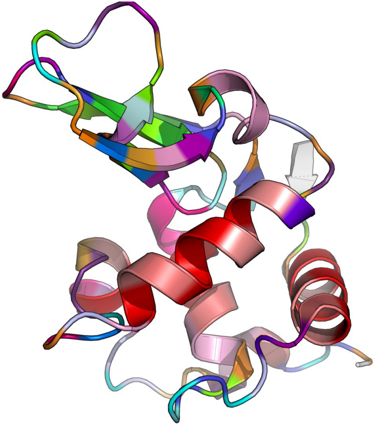
- Camproux et al., *Protein Eng.*, 1999.
- Camproux et al., *J Mol Biol.*, 2004.
- Camproux and Tufféry, *Biochim Biophys Acta.*, 2005.

Structural Alphabet (SA)

An encoding example: 135L

```
>Amino Acids
KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNT
HATNRNTDGSTDYGILQINSRWWCNDGRTPGSKNLCNIPC
SALLSSDITASVNCAKKIASGGNGMNAWVAWRNRCKGTDV
HAWIRGCRL
```

```
>HMMSA
NLHWAAAAAVWAVDOQUSUFSLHBBVWAAAVZZFFFSPBS
XTLNHZDSNLNJFZDRLPECCILGDEQLUGPRGBDSKHBB
BBQHEGOWAVWAAAVWABQHZRUEEEGWAAZCCQUQYGEB
BVSUSP
```



Outline

1. HMM based Structural Alphabet
   - HMM-SA27
   - Structural Alphabet (SA)
2. SAFrAN
   - SA prediction from amino acids sequence
   - SA Search
   - SAFrAN algorithm
   - SAFrAN example
   - Candidate fragments properties
3. Greedy-OPEP
   - Greedy algorithm
   - OPEP: a coarse-grain force field
4. Results - Discussion
   - CASP7 experiment
   - Improvements since CASP7

SA prediction from amino acids sequence

HMM-SA / Amino acids dependency

$$p(AA_i/SA_i) \tag{1}$$

Process learnt from a non redundant collection of HMM-SA encoded proteins.

$\hookrightarrow$ *Constrain prediction on a subset of HMM-SA letters*

SA prediction from amino acids sequence

HMM-SA / Amino acids dependency

$$p(AA_i/SA_i) \tag{1}$$

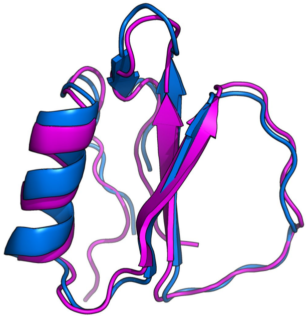
Process learnt from a non redundant collection of HMM-SA encoded proteins.

↪ *Constrain prediction on a subset of HMM-SA letters*

- Use **PSIPRED** (Jones D., *J Mol Biol.*, 1999)
- Confidence level threshold: **5** (min: 0 / max: 9)
  - **helices**: **(a A V W Z B C D E)**,
  - **strands**: **(L M N T X J K)**,
  - **coils**: **(B C D E F G H I J K L N O P Q R S T U Y Z)**,
  - others: the full alphabet is used (27 letters).

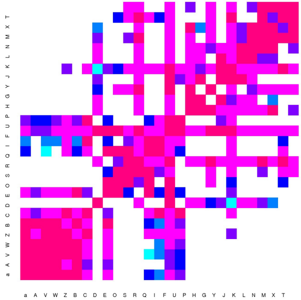
SA Search

```
>P1;2ci2I [a]
MGBEOUSKHVWVWAVWAAZCGZSMNTMXKKUSLNKHSLT
PZQTMNMN-MYZDSKGIYLXK*
TEWPELVGKSVEEAKKVILQDKPEAQIIVLPVGTIVTME
YRIDRVRL-FVDKLDNIAEVP*
>P1;1cseI
MGBBQUSKHAAVWVWVWZCCGBQPRNMXKKUSKXYPQKT
PVQTMNNXKLUaDSPGQKLNK*
KSFPEVVGKTVDQAREYFTLHYPQYNVVYFLPEGSPVTLD
LRYNRVRVFYNPGTNVVNHVP*
```



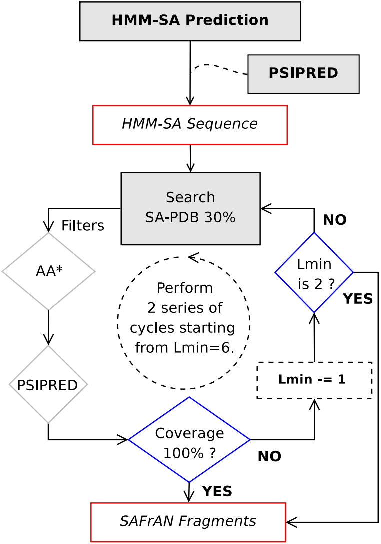
Search for structural similarities

- 3D structures could be **aligned in HMM-SA space**.
- **Exact matches** (*Suffix tree*)
- **Fuzzy matches** (*Dynamic programming*)
  ↪ **Substitution Matrix**



[a] HMMSA / AA alignments

SAFrAN algorithm



* AA Filter evolves during series of cycles.

SAFrAN steps are:

1. **Predict HMM-SA sequence from amino acid sequence**, conditionnaly to PSIPRED.
2. **Search for compatible words** in a non redundant PDB with classical alignment tools (*Smith and Waterman*).
3. **Filter solutions** (Amino acids sequences compatibility, PSIPRED compatibility and redundancy).
4. **Decrease the minimal match length**.
5. **Iterate** until the full coverage of the sequence or no more words could be reached.

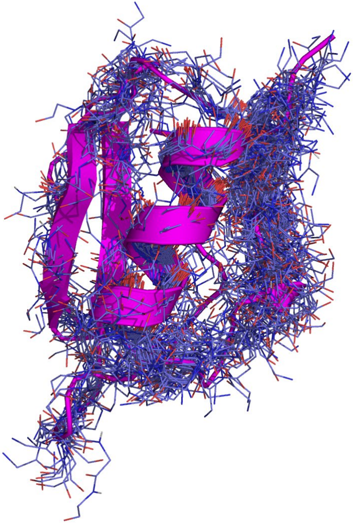
SAFrAN example

SAFrAN typical output (HMM-SA)

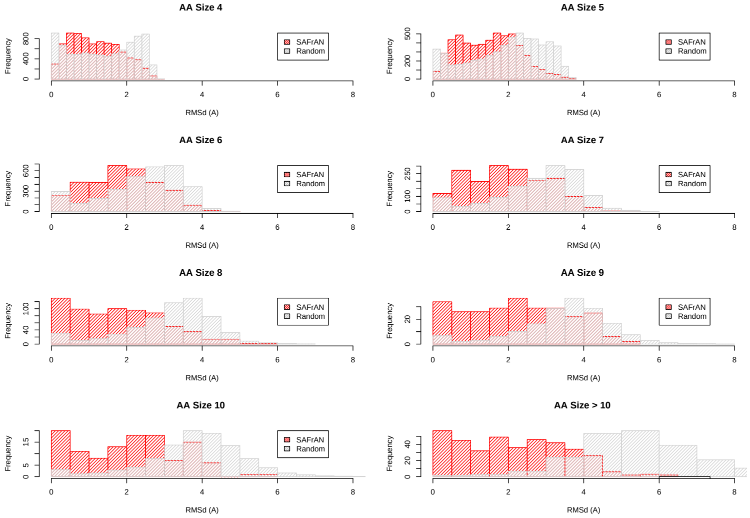
```
MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYA    ### QUERY ###
ITSNL-----------------------------     1     5  2ez9A   406   410
VSWQLN----------------------------     1     6  1u7iA   126   131
[...]
-TAGIIVAG-------------------------     2     9  1rypH   103   110
--LRVVFSG-------------------------     3     9  1xk7A    17    23
[...]
----LKLFGESI----------------------     5    12  1r64A   468   475
------RSGRITL---------------------     7    13  1musA   263   269
-------DGLIIPGL-------------------     8    15  1njrA    52    59
[...]
-----------FEGTTT-----------------    12    17  1czfA    47    52
[...]
-------------GVRTAEDAQKYLAIADELF--    14    32  1p1xA   205   223
--------------GTQREHIDLANACKEIFIKE    15    34  2cfaA    63    82
[...]
------------------------EALKAFHELS    25    34  1v8zA   326   335
[...]
-------------------------------RFA    32    44  1xdnA    56    59
```

Matching fragments

superposed on the target structure **2CI2**.



Candidate fragments properties



Outline

1. HMM based Structural Alphabet
   - HMM-SA27
   - Structural Alphabet (SA)

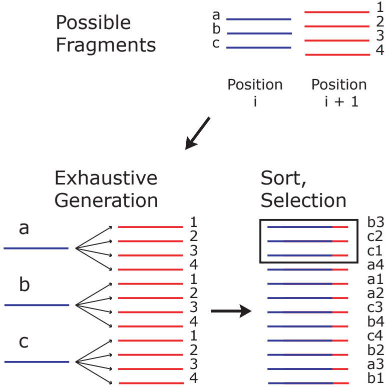
2. SAFrAN
   - SA prediction from amino acids sequence
   - SA Search
   - SAFrAN algorithm
   - SAFrAN example
   - Candidate fragments properties

3. Greedy-OPEP
   - Greedy algorithm
   - OPEP: a coarse-grain force field

4. Results - Discussion
   - CASP7 experiment
   - Improvements since CASP7

Greedy algorithm

The original greedy algorithm



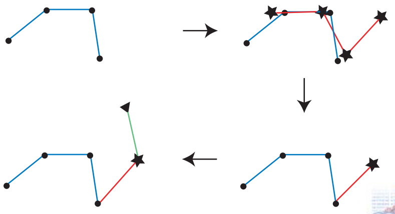
Inspired from Kolodny et al., *J Mol Biol.*, 2002.

- Tuffery et al., *J Comput Chem.*, 2005
- Tuffery and Derreumaux, *Proteins.*, 2005

Greedy algorithm improvements

- Prefiltering
- Iterated
- Randomized

Superposition procedure:

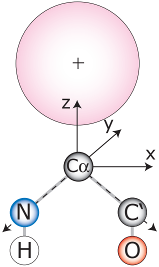


OPEP: a coarse-grain force field

Optimized Potential for Efficient peptide structure Prediction

$$E_{OPEP} = E_{SC,SC} + E_{C\alpha,C\alpha} + E_{VdW} + E_{HB} + \\ E_{bonds} + E_{angles} + E_{imp-torsions} + E_{\phi>0} \quad (2)$$

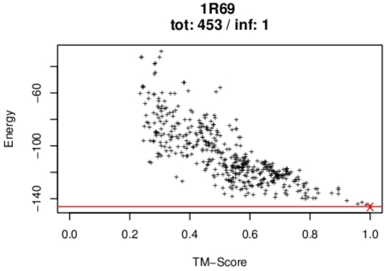
6-bead model

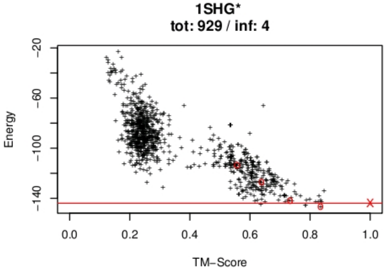


- N, HN, Cα, C, O atoms are explicit.
- Side Chains are represented by one bead.

Santini et al., *Internet Electron. J. Mol. Des.*, 2003.

OPEP: a coarse-grain force field



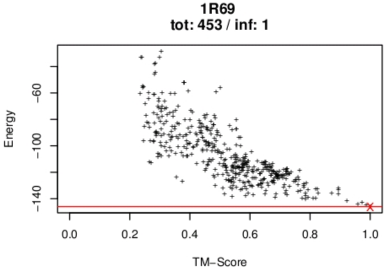


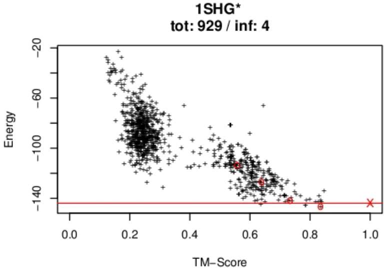
OPEP Optimisation

- Trained and validated on generated and publicly available decoys sets.
- OPEP is able to **find a native like structure for 24 targets on 29** of our decoys sets.

Maupetit et al., *Proteins.*, 2007.

OPEP: a coarse-grain force field


1R69
tot: 453 / inf: 1


1SHG*
tot: 929 / inf: 4

Maupetit et al., *Proteins.*, 2007.

OPEP Optimisation

- Trained and validated on generated and publicly available decoys sets.
- OPEP is able to **find a native like structure for 24 targets on 29** of our decoys sets.

sOPEP Greedy implementation

$$E_{sOPEP} = E_{SC,SC} + E_{C\alpha,C\alpha} + E_{VdW} + E_{HB} + E_{\phi>0} \quad (3)$$

Outline

1. HMM based Structural Alphabet
   - HMM-SA27
   - Structural Alphabet (SA)
2. SAFrAN
   - SA prediction from amino acids sequence
   - SA Search
   - SAFrAN algorithm
   - SAFrAN example
   - Candidate fragments properties
3. Greedy-OPEP
   - Greedy algorithm
   - OPEP: a coarse-grain force field
4. Results - Discussion
   - CASP7 experiment
   - Improvements since CASP7

CASP7 experiment

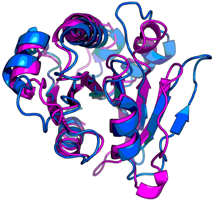
SAFrAN performances.

- The major part of the sequence is covered (94%)
- SAFrAN derived **HMM-SA trajectories** could lead to **near native** solutions (< 2.0 Å).
- The **complexity**, *ie* average number of prototypes used at each HMM-SA position, is 13 when using 3 prototypes maximum by HMM-SA letter.

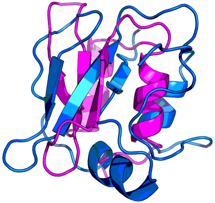
CASP7 experiment

Greedy performances.

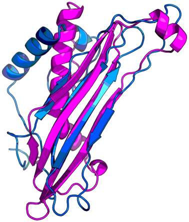
t0308 (HA-TBM)



t0358 (FM)



t0383 (FM)

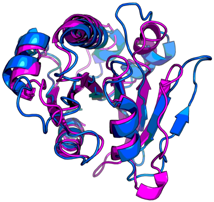


Native / Model.

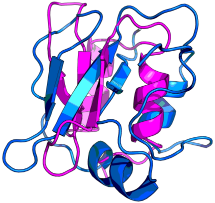
CASP7 experiment

Greedy performances.

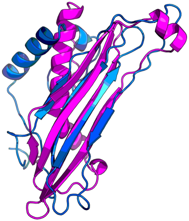
t0308 (HA-TBM)



t0358 (FM)



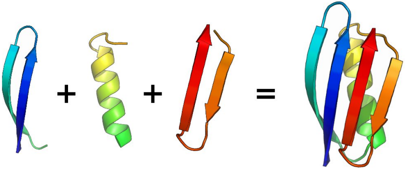
t0383 (FM)



Native / Model.

- **Hierarchical approach** leads to best results.

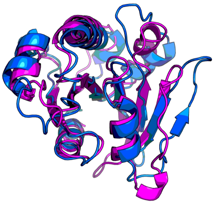
An example of hiearchical approach.



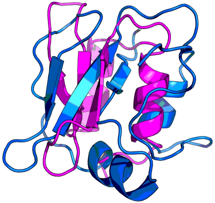
CASP7 experiment

Greedy performances.

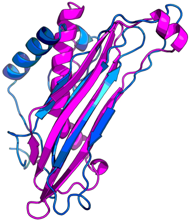
t0308 (HA-TBM)



t0358 (FM)



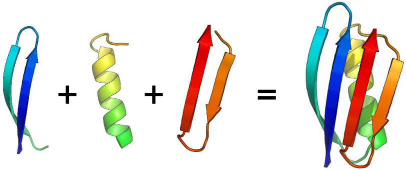
t0383 (FM)



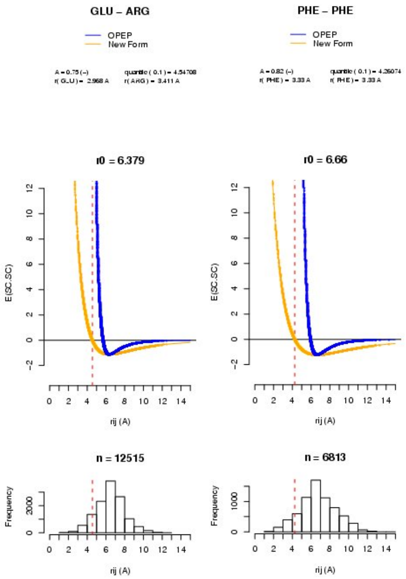
Native / Model.

- **Hierarchical approach** leads to best results.
- **Side chains interactions** were not optimal for our discrete assembling procedure.

An example of hiearchical approach.



Side chains interactions improvements

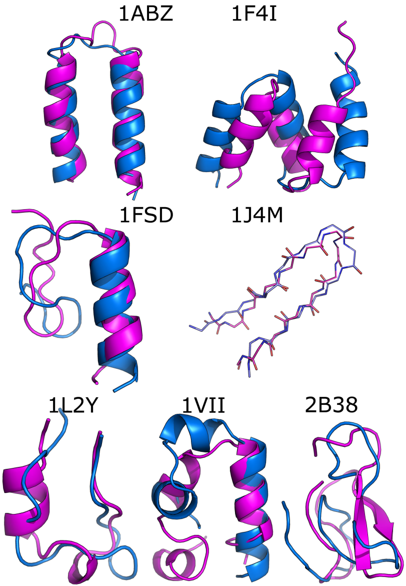


New formulation

Find parameters that best fit the interacting centroids distance distribution.

- Smooth the potential
- Lowest energy for the mean distance
- Start to penalize interaction for a quantile of 10%.

To a hierarchical approach ?

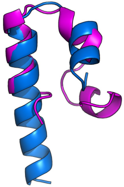


Hierarchical approach

- Are we able to build small peptides correctly ?
- Best results combined with the new PMF formulation.

Mean RMSd: **3.9 Å** (*vs* 4.7 Å).

1VII with OPEP v3 PMF formulation.



Conclusions & perspectives

Conclusions:

- **SAFrAN**′s method gives **promising results**.
- SAFrAN could be useful to assist structure resolution from experimental data.

Perspectives:

- **Homologous protein detection**: SAFrAN ?
- **Hierarchical procedure**: how to split protein structures into supersecondary structure elements ?
- **sOPEP force field improvements ?** Are OPEP parameters optimal for a discrete modeling procedure ?
- Complete method **automatization**.

Have contributed to this work

EBGM, Paris.

- Pierre Tufféry *(HMMSA, SAFrAN, Greedy, OPEP, sOPEP)*
- Frédéric Guyon *(HMMSA, SAFrAN, Greedy)*
- Anne-Claude Camproux *(HMMSA, SAFrAN)*

INSERM U726, Université Paris Diderot.

IBPC, Paris.

- Philippe Derreumaux *(Greedy, OPEP)*

CNRS UPR 9080, Université Paris Diderot.

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Inserm
Institut national
de la santé et de la recherche médicale

université PARIS DIDEROT PARIS 7

	FM	**HA-TBM**	**TBM**	**TOT**
# targets	4	9	2	**15**
% Coverage	97%	94%	88%	**94%**
Search complexity (1)	23.43 ±5.20	19.43 ±6.8	18.86 ±4.74	**20.49** ±6.05
Search complexity (2)	14.11 ±2.61	12.19 ±3.51	12.31 ±2.02	**12.72** ±3.09
Best Rebuilt cRMSd (1)	0.88 Å ±0.49	1.62 Å ±0.52	1.34 Å ±0.18	**1.39 Å** ±0.57
Best Rebuilt cRMSd (2)	1.12 Å ±0.41	1.96 Å ±0.54	1.68 Å ±0.33	**1.70 Å** ±0.59

(1) Using all prototypes by letter. (2) Using 3 prototypes maximum by letter. TOT = FM + HA-TBM + TBM.

SAFrAN performances.

- The quite full sequence is covered (94%)
- SAFrAN derived **HMM-SA trajectories** could lead to **near native** solutions (< 2.0 Å).
- The **complexity**, *ie* average number of prototypes used at each HMM-SA position, is 13 when using 3 prototypes maximum by HMM-SA letter.