ASSESSING A NEW APPROACH FOR PROTEIN STRUCTURE MODELING COMBINING STRUCTURAL ALPHABET LOCAL CONFORMATION PREDICTION AND GREEDY ALGORITHM FOR RECONSTRUCTION.

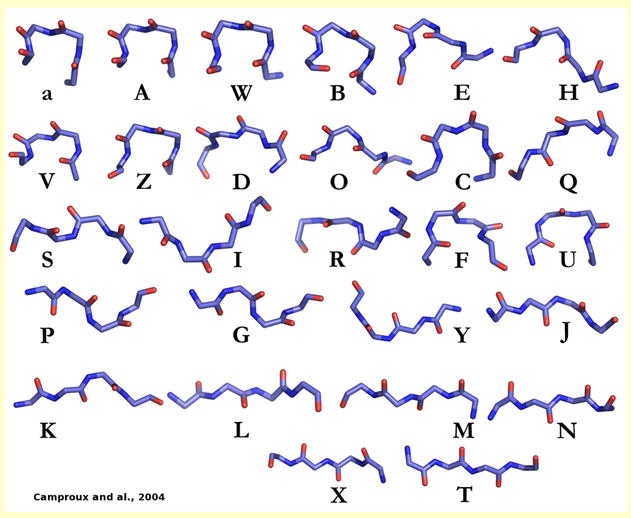
J. Maupetit[1], F. Guyon[1], J. Martin[1], A.C. Camproux[1], Ph. Derreumaux[2] and Pierre Tufféry[1]

1 *Equipe de Bioinformatique Génomique et Moléculaire,* INSERM E0346, Université Paris 7, Tour 53/54 1er Etage, 2 place Jussieu, 75251 Paris Cedex 05, France

2 *Laboratoire de Biochimie Théorique,* UPR 9080 CNRS, IBPC et Université Paris 7, 13 rue Pierre et Marie Curie, 75005 Paris, France

I. MATERIALS AND METHODS:

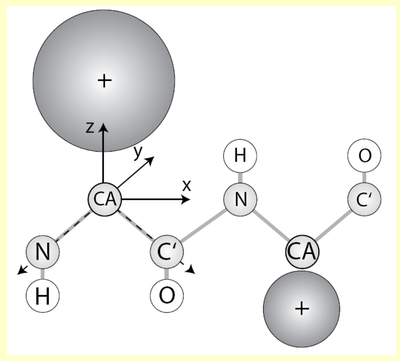
***HMM-SA* [1]:** To describe protein local conformation, we use a Hidden Markov Model (HMM) learn from 1429 PDB structures. Each letter of the alphabet is a 4 residues length protein fragment. Consecutive fragments overlapp by 3 residues. We use a 27 letters structural alphabet. Each letter is associated with a canonical conformation. Since we perform rigid discrete assembly, we allow several sub-conformation per letter (a total of 155 prototypes to describe the 27 letters).



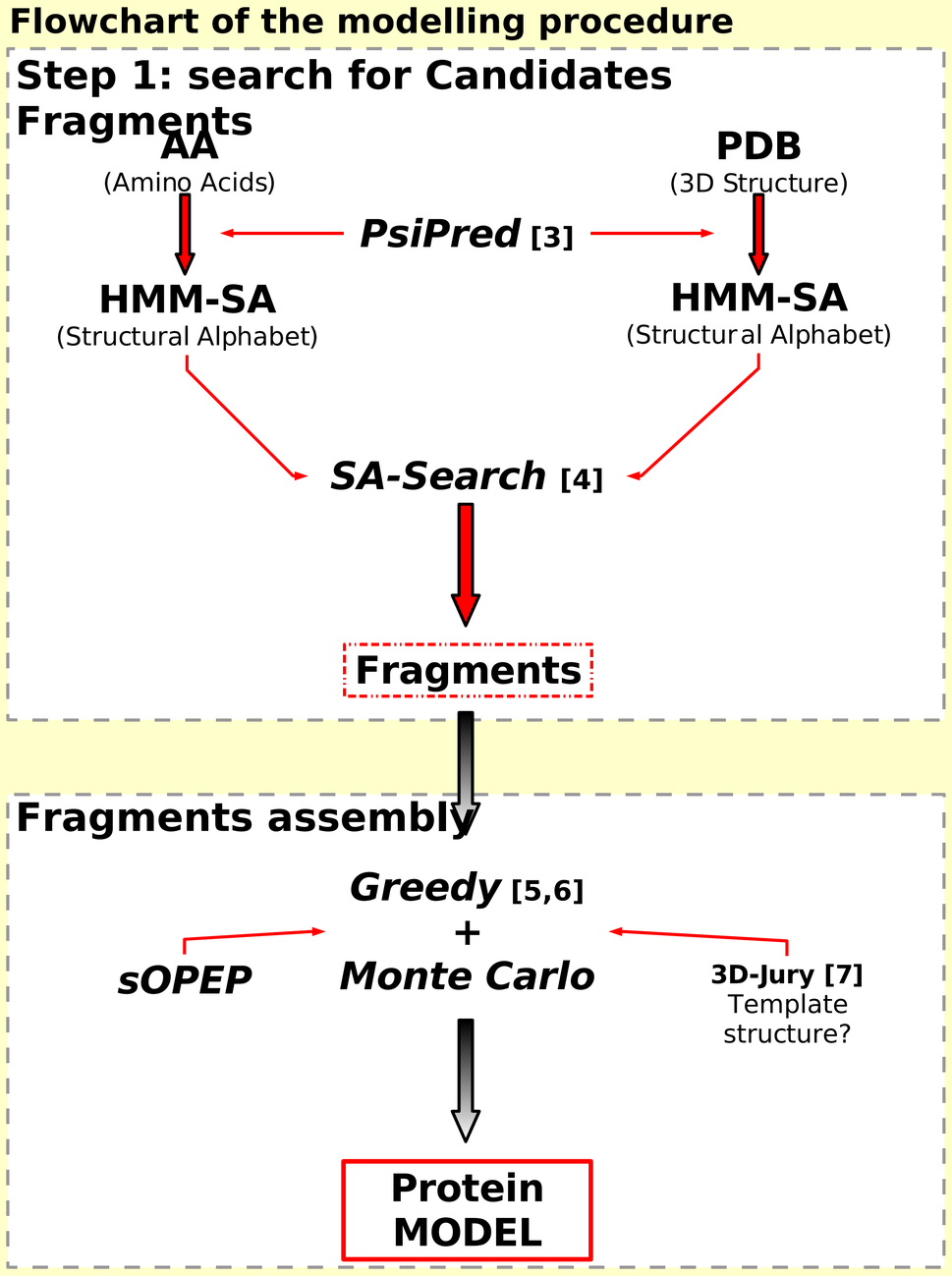
***sOPEP* :** a simplified version of OPEP **[2]** is used to drive model generation :

$$E_{sOPEP} = E_{VdW} + E_{C_\alpha C_\alpha} + E_{PMF} + E_{\Phi>0} + E_{HB}$$

Coarse grained sOPEP representation: Main chain is explicit and side chains are represented by one bead with a diameter depending on the considered amino acid.

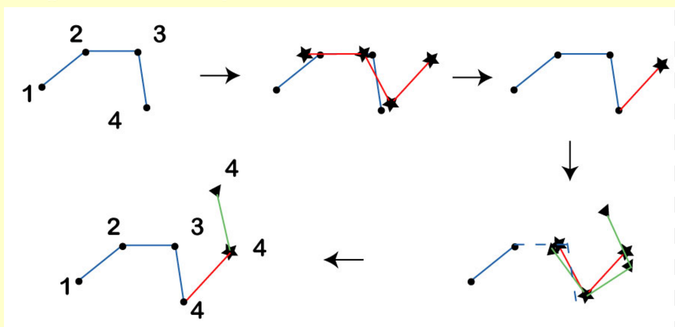


Flowchart of the modelling procedure

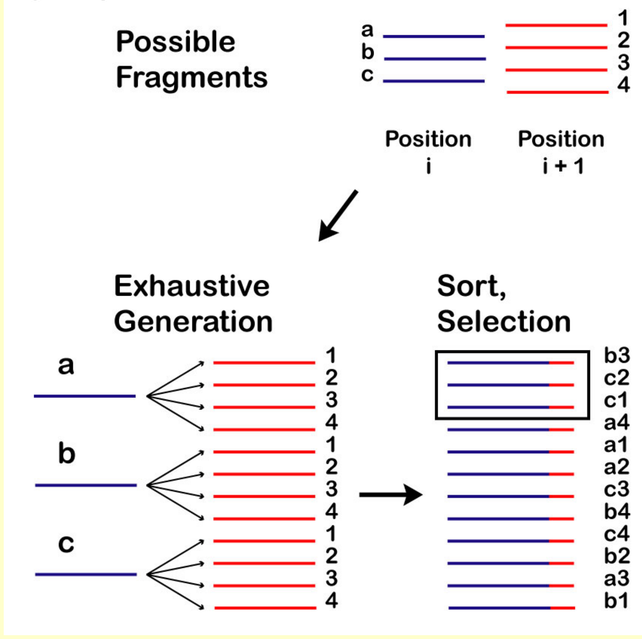


« Greedy » fragment assembly

Fragments overlap:



The greedy procedure:



Final refinement: once CA trace is determined, full atom models are generated by SABBAC **[8]**, and then minimized by the GROMACS **[9]** package.

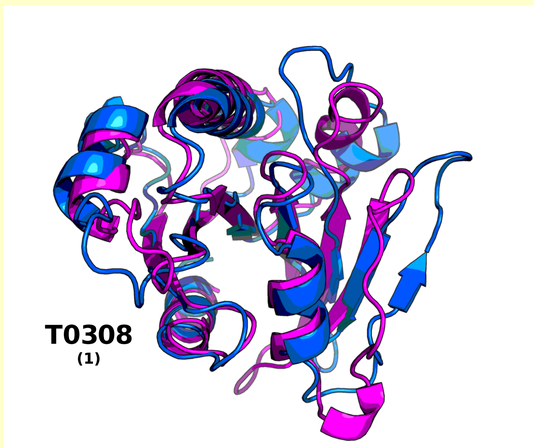
II. ANALYSING CASP7 RESULTS:

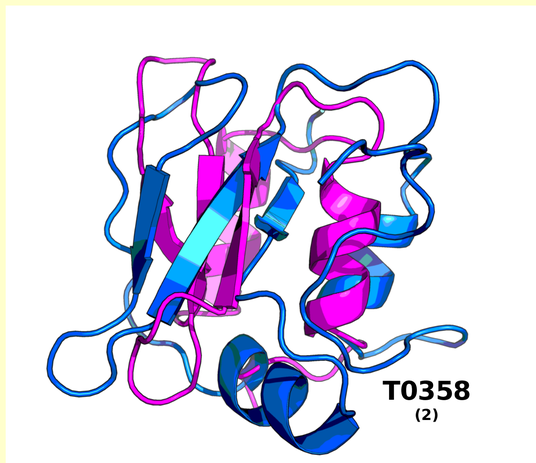
1. Assessing candidate fragments quality

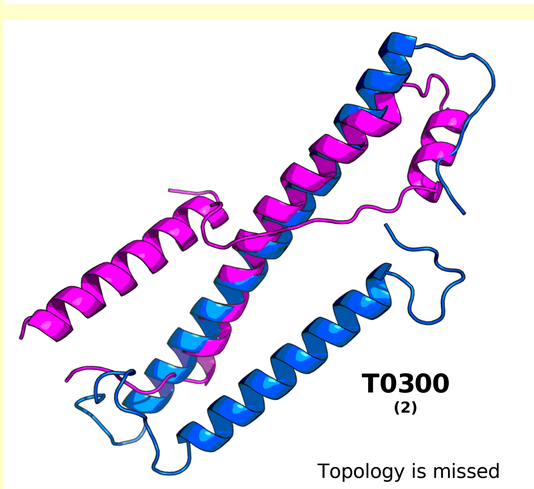
	CutPref (#4)	*PDB Blast (#9)*	*3D Jury (#2)*	TOT
% Coverage	**97**%	**94**%	**88**%	**94**%
Search complexity *All Prototypes*	**23.43** ±5.20	**19.43** ±6.8	**18.86** ±4.74	**20.49** ±6.05
Search complexity *Max 3 prototypes by letter*	**14.11** ±2.61	**12.19** ±3.51	**12.31** ±2.02	**12.72** ±3.09
# HMM-SA letters per Pos	**7.44** ±4.84	**5.99** ±4.82	**5.61** ±4.56	**6.20** ±4.83
Best Rebuilt RMSd *All Prototypes*	**0.88 A** ±0.49	**1.62 A** ±0.52	**1.34 A** ±0.18	**1.39 A** ±0.57
Best Rebuilt RMSd *Max 3 prototypes by letter*	**1.12 A** ±0.41	**1.96 A** ±0.54	**1.68 A** ±0.33	**1.70 A** ±0.59

% Coverage: fraction of the protein described by candidate fragments. Remaining parts are filled using direct HMM-SA prediction from sequence.
Search complexity: Average number of rigid fragment used per residue during model generation.
 # HMM-SA letters per Pos: Average number of SA letters describing each position (max is 27 - means everything, i.e. no prediciton).
Best build RMSd: We use the native structure to assess of accurately the candidate fragments can reconstruct it. This is our best possible reconstruction accuracy.
Targets are classified according to the Robetta server [11] classification (Cutpref, 3DJury, pdblast)

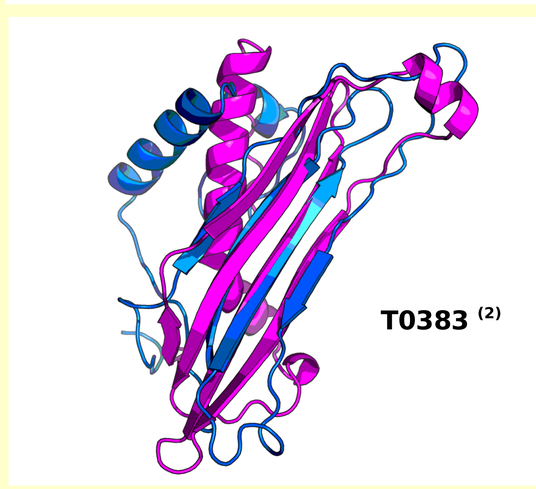
2. Assessment of model generation procedure







Topology is missed



Some CASP7 models: in all cases, the native structure is colored in magenta, and our superposed model in marine blue. (1) “pdbblast” targets (2) “cutpref” targets. Pictures generated using the PyMol software **[10]**.

III. Conclusions, perspectives:

We have taken the opportunity of CASP7 to assess a new model generation procedure. Only few targets were submitted due to concurrent developpment / improvement of the procedure during CASP. Emphase was put on potentially difficult target.

Coming out:
- Candidate fragment selection seems efficient. The solution is in the selected fragments.
- Complete procedure, starting from sequence only (no template, de novo model generation) was able to produce in some case topologically satisfactory models.

Limitations:
- Assembly is too rigid for homology modelling
- The procedure needs a final model refinement able to shake the whole structure.

References:

[1] Camproux AC, Gautier R, Tuffery P. A hidden markov model derived structural alphabet for proteins. *J Mol Biol.* 2004 Jun 4;339(3):591605.
[2] P. Derreumaux, Generating ensemble averages for small proteins from extended conformations by Monte Carlo simulations. Phys. Rev. Lett. (2000) 85: 206-209.
[3] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 1999 Sep 17;292(2):195-202.
[4] Guyon F, Camproux AC, Hochez J, Tuffery P. SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Res.* 2004 Jul 1;32(Web Server issue):W545-8.
[5] Tuffery P, Guyon F, Derreumaux P. Improved greedy algorithm for protein structure reconstruction. *J Comput Chem.* 2005 Apr 15;26(5):50613.
[6] Tuffery P, Derreumaux P. Dependency between consecutive local conformations helps assemble protein structures from secondary structures using Go potential and greedy algorithm. *Proteins.* 2005 Dec 1;61(4):73240.
[7] Ginalski K, Elofsson A, Fischer D, Rychlewski L "3D-Jury: a simple approach to improve protein structure predictions." *Bioinformatics.*(2003 **19**(8):1015-8.
[8] Maupetit J, Gautier R, Tuffery P. SABBAC: online Structural Alphabet based protein BackBone reconstruction from AlphaCarbon trace. *Nucleic Acids Res.* 2006 Jul 1;34(Web Server issue):W14751.
[9] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ GROMACS: fast, flexible, and free. *J Comput Chem.* (2005) **26**(16):1701-18.
[10] DeLano, W.L. The PyMOL Molecular Graphics System (2002) *DeLano Scientific, San Carlos, CA, USA.* http://www.pymol.org
[11] Kim DE*, Chivian D*, Baker D. (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 32 Suppl 2:W526-31