A Greedy Algorithm for Protein Structure Reconstruction:

Improvements and Applications

Julien Maupetit¹, Philippe Derreumaux² and Pierre Tufféry³

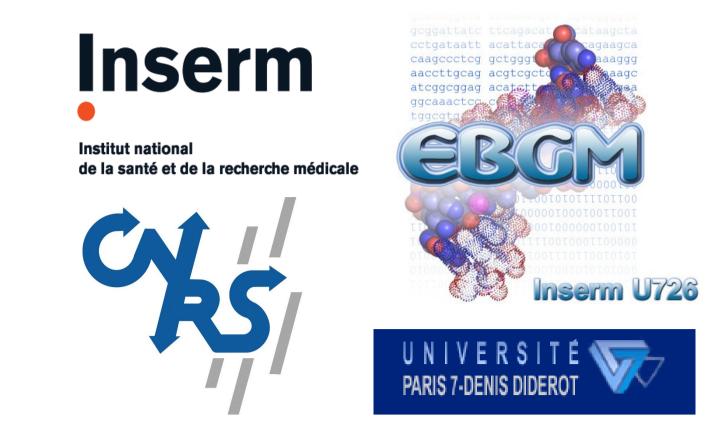
1,3 Equipe de Bioinformatique Génomique et Moléculaire, INSERM E0346, Université Paris 7, Tour 53/54 1er Etage, 2 place Jussieu, 75251 Paris Cedex 05, France

Goal: Large scale genome sequencing projects bring sequence databases to grow exponentially. Even if the number of known proteins structures increases rapidly, only few of the detected ORFs correspond to proteins having an experimentally resolved 3D structure. For such reason, one needs to develop high throughput methods to model 3D protein structures. Moreover, it is important that modeling methods also be able to generate structures for the proteins that fall out the range application of comparative modeling techniques.

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

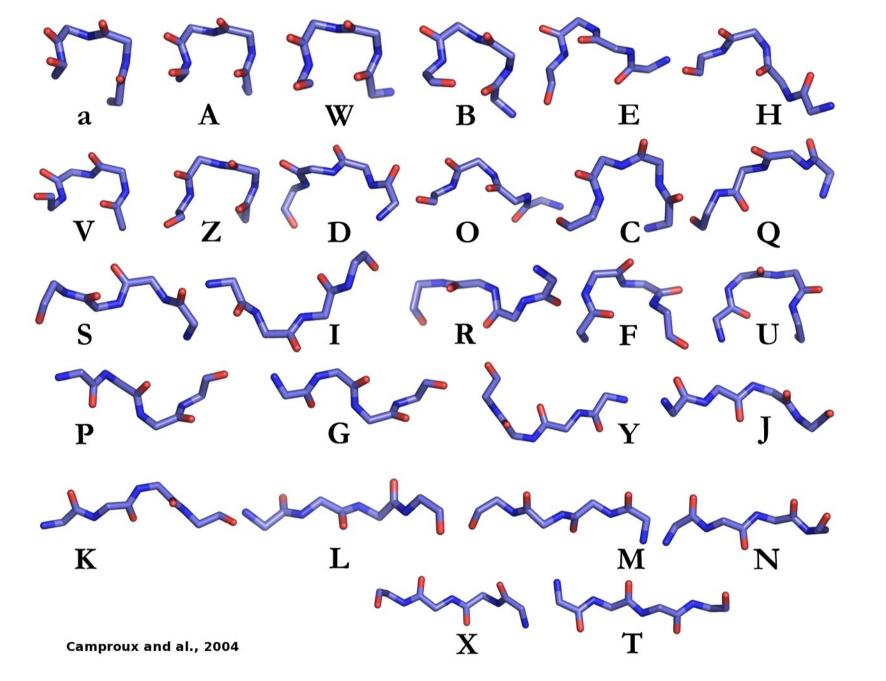
Objective function: to test the capability of the algorithm to reconstruct protein structures, alpha carbons RMSd (cRMSd) criterion was first used. It has been shown that greedy algorithm is able to rebuild protein structures at less than 1A cRMSd [2].

In a second time, an improved version of the greedy algorithm (see below) in combination with a Go based criterion, lead to structures differing by less than 4.8A cRMSd [**2**].



SABBAC (Structural Alphabet based BackBone reconstruction from Alpha-Carbon trace) [5]: based on our greedy algorithm, we implemented a web tool to rebuild protein backbone from alpha-carbon trace. Alphacarbon coordinates remain unaffected. SABBAC simply positions the missing backbone atoms with an energetic criteria, no further refinement is performed. From our tests, SABBAC performs equal or better than other similar on-line approach and is robust to deviations of the alpha carbons coordinates. It is available at

Structural alphabet (SA) [1]: the present method is based on a Hidden Markov Model (HMM) learn from 1429 PDB structures. Each letter of the alphabet is a 4 residues length protein fragment, overlapping on 3 residues. HMM is described by 4 distances : C1C3, C1C4, C2C4, and the projection of C4 on the plane described by C1C2C3. This results in 27 states or letters (155 prototypes).

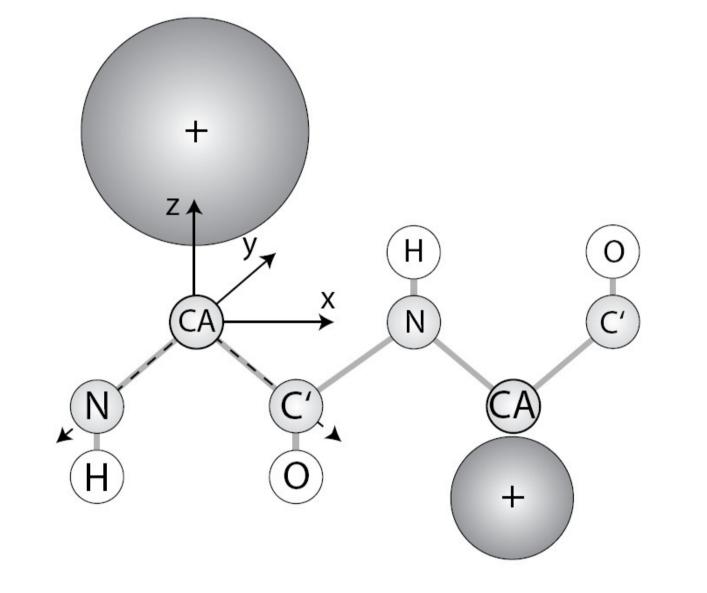


HMM-SA description Generate the of **model:** there are two different ways to determine the collection of fragments to use at each position of the protein : prediction or assignation. Prediction starts from the amino acid sequence to propose structural alphabet letters. Assignation starts from known 3D structure, and searches for fragments best describing short parts of the structure.

$$E_{\rm Go} = \sum_{j>i} \Delta_{ij} B_{ij} + E_{\rm reg}$$

Finally, in an *ab initio* context, we started to implement a simplified version of the coarse grained OPEP [4] force field (*sOPEP*) with the following energetic terms:

$$\mathcal{E}_{sOPEP} = \mathcal{E}_{VdW} + \mathcal{E}_{C_{\alpha}C_{\alpha}} + \mathcal{E}_{PMF} + \mathcal{E}_{\Phi>0} + \mathcal{E}_{H}$$



http://bioserv.rpbs.jussieu.fr/cgi-bin/SABBAC

B

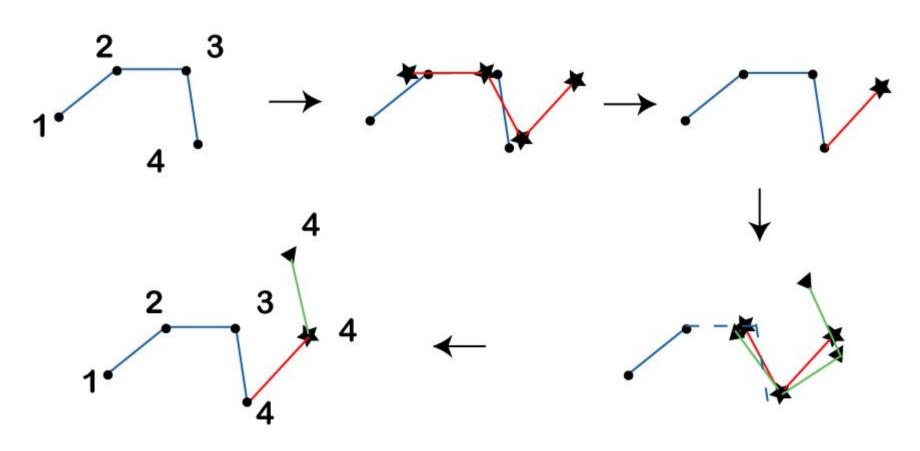
PDB	n Residues	Main chain RMSd			▼ SABBAC - Mozilla Firefox Fichier Editon Affichage Aljerà Marque-pages Ouțis Ajde		
		SABBAC	MaxSprout	bb			
Adcock subset							
111M	154	0.29	0.42	0.91	SABBAC: Structural Alphabet based protein Backbone Builder from Alpha Carbon trace. This tool is for the reconstruction of complete protein structures from their alpha-carbon description. Side chains positionning is performed by SCI methods. You have to choose between: Upload a PDB file with CA coordinates Paste PDB file to text area in from text area in from text area. Type a PDB file with CA coordinates Paste a PDB URL instead of an id in the text area. If you specify both upload and base data and the upload will be considered Note text: SABBAC currently only accepts files having one chain. SABBAC does not accept presently non standard amino acids. More information here Image: Part of PDB File		
1CTF	68	0.43	0.73	0.85			
1IGD	61	0.36	0.44	0.68			
10MD	107	0.35	0.41	0.77			
1SEMA	58	0.48	0.34	1.00			
1TIMA	247	0.59	0.60	0.97			
1UBQ	76	0.35	0.38	0.96			
2CTS	437	0.40	0.38	0.96			
	10-1						
2LYM	129	0.38	0.44	0.98			
2MHR	118	0.50	0.54	0.88			
2PCY	99	0.42	0.54	0.91			
2WRP	104	0.30	0.42	0.87			
4PTI	58	0.53	0.44	0.81			
5NLL	138	0.37	0.46	0.85	Process Clear		
Mean		0.41	0.47	0.89	III Terminé		
SD		0.09	0.10	0.08			
DB newcome	ers subset						
DD newconne							
1PXZA	346	0.55	0.54	0.96			
	346 101	0.55 0.58	0.54 0.44	0.96 0.88			
1PXZA					✓ SABBAC Results - Mozilla Firefox Fichier Editon Affichage Alferà Manue-pages Outils Ajde		
1PXZA 1RKIA	101	0.58	0.44	0.88			
1PXZA 1RKIA 1S7LA	101 177	0.58 0.29	0.44 0.36	0.88 0.86	Fichier Edition Affichage Aller & Marque-pages Ouglis Alde Image: Page Aller & Marque-pages Ouglis Alde Image: Page Aller & Marque-pages Ouglis Alde Image: Page Aller & Marque-pages Ouglis Aller Image: Page Aller Image: Page Aller		
1PXZA 1RKIA 1S7LA 1T70A	101 177 255	0.58 0.29 0.42	0.44 0.36 0.50	0.88 0.86 0.95	Fichier Edition Affichage Alferà Marque-pages Ouțils Alde Inter/Noiserv.rpbs.jussieu.fr/mp/SABBAC/8139262/index.html O OK (8)		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA	101 177 255 235	0.58 0.29 0.42 0.41	0.44 0.36 0.50 0.38	0.88 0.86 0.95 0.96	Fichier Edition Affichage Aller & Marque-pages Ouglis Alde Image: Aller & Marque-pages Ouglis Alde Image: Aller & Marque-pages Ouglis Alde Image: Aller & Marque-pages Ouglis Aller		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED	101 177 255 235 666	0.58 0.29 0.42 0.41 0.48	0.44 0.36 0.50 0.38 0.45	0.88 0.86 0.95 0.96 0.89	Ficher Editor Affichage Aller à Marque-pages Ouțils Alde Image: Alger à Marque-pages Ouțils Alde Image: Alger à Marque-pages Ouțils Alger Image: Alger à Marque-pages Outils Alger		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA	101 177 255 235 666 175	0.58 0.29 0.42 0.41 0.48 0.30	0.44 0.36 0.50 0.38 0.45 0.41	0.88 0.86 0.95 0.96 0.89 0.87	Fichier Edition Affichage Aller & Marque-pages Ouglis Alde Image: Aller & Marque-pages Ouglis Alde Image: Aller & Marque-pages Ouglis Alde Image: Aller & Marque-pages Ouglis Aller		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B	101 177 255 235 666 175 255	0.58 0.29 0.42 0.41 0.48 0.30 0.34	0.44 0.36 0.50 0.38 0.45 0.41 0.42	0.88 0.86 0.95 0.96 0.89 0.87 0.84	Ficher Editor Allerà Marque-pages Ouțils Aide Image: Section Allerà Section All		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A	101 177 255 235 666 175 255 149 103	$\begin{array}{c} 0.58 \\ 0.29 \\ 0.42 \\ 0.41 \\ 0.48 \\ 0.30 \\ 0.34 \\ 0.28 \\ 0.47 \end{array}$	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82	Ficher Editor Allerà Marque-pages Ouțils Aide Image: Section Allerà Section All		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A 1VR9A	101 177 255 235 666 175 255 149	0.58 0.29 0.42 0.41 0.48 0.30 0.34 0.28 0.47 0.42	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59 0.45	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82 1.00 0.79	Ficher Edition Affichage Allerà Marque-pages Ouglis Alde Image: Allerà Marque-pages Ouglis Allerà Marqu		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A 1VR4A 1VR9A 1WMHA	101 177 255 235 666 175 255 149 103 121 83	$\begin{array}{c} 0.58 \\ 0.29 \\ 0.42 \\ 0.41 \\ 0.48 \\ 0.30 \\ 0.34 \\ 0.28 \\ 0.47 \\ 0.42 \\ 0.27 \end{array}$	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59 0.45 0.28	$\begin{array}{c} 0.88\\ 0.86\\ 0.95\\ 0.96\\ 0.89\\ 0.87\\ 0.84\\ 0.82\\ 1.00\\ 0.79\\ 0.82\end{array}$	Ficher Edition Affichage Aller à Marque-pages Ouglis Aide Image: Aller à Marque-pages Ouglis Aide Image: Aller à Marque-pages Ouglis Aller		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A 1VR4A 1VR9A 1WMHA 1WPBG	101 177 255 235 666 175 255 149 103 121 83 168	0.58 0.29 0.42 0.41 0.48 0.30 0.34 0.28 0.47 0.42 0.27 0.27 0.37	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59 0.45 0.28 0.35	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82 1.00 0.79 0.82 0.86	Ficher Edition & Michage Aller & Manue-pages Outlis Aide Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Image: Section & Michage Aller & Manue-pages Outlis Aller Results for your request. Submit another job. Results will be available for 5 days PDB reference: Image: directions Image: Michage Aller & Manue-pages Page: Michage Aller & Michage Image: Michage Aller		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR5B 1VKCA 1VR4A 1VR9A 1WR9A 1WMHA 1WPBG 1WMIA	101 177 255 235 666 175 255 149 103 121 83 168 88	$\begin{array}{c} 0.58 \\ 0.29 \\ 0.42 \\ 0.41 \\ 0.48 \\ 0.30 \\ 0.34 \\ 0.28 \\ 0.47 \\ 0.42 \\ 0.27 \\ 0.37 \\ 0.41 \end{array}$	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59 0.45 0.28 0.35 0.42	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82 1.00 0.79 0.82	Ficher Edition Affichage Aller à Marque-pages Ouglis Aide Image: Aller à Marque-pages Ouglis Aide Image: Aller à Marque-pages Ouglis Aller		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A 1VR9A 1WR4A 1WR9A 1WMHA 1WPBG 1WMIA 1X6JA	101 177 255 235 666 175 255 149 103 121 83 168 88 88	$\begin{array}{c} 0.58 \\ 0.29 \\ 0.42 \\ 0.41 \\ 0.48 \\ 0.30 \\ 0.34 \\ 0.28 \\ 0.47 \\ 0.42 \\ 0.27 \\ 0.42 \\ 0.27 \\ 0.37 \\ 0.41 \\ 0.43 \end{array}$	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59 0.45 0.28 0.35 0.42 0.36	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82 1.00 0.79 0.82 0.82 0.86 0.81 0.76	Ficher Edition Affichage Aller & Marque-pages Ouglis Aide Image: Aller & Marque-pages Ouglis Alled Image: Aller & Marque-pages Ouglis Aller		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A 1VR9A 1WR4A 1WR9A 1WMHA 1WPBG 1WMIA 1X6JA 1XB9A	101 177 255 235 666 175 255 149 103 121 83 168 88 88 88 108	$\begin{array}{c} 0.58\\ 0.29\\ 0.42\\ 0.41\\ 0.48\\ 0.30\\ 0.34\\ 0.28\\ 0.47\\ 0.42\\ 0.27\\ 0.37\\ 0.41\\ 0.43\\ 0.46\end{array}$	$\begin{array}{c} 0.44\\ 0.36\\ 0.50\\ 0.38\\ 0.45\\ 0.41\\ 0.42\\ 0.33\\ 0.59\\ 0.45\\ 0.28\\ 0.35\\ 0.42\\ 0.36\\ 0.51\\ \end{array}$	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82 1.00 0.79 0.82 0.82 0.86 0.81 0.76 0.81	Ficher Edition Alleria Marque-pages Ouglis Adde Image: Section Marque-pages Ouglis Ouglis Marque-pages Ouglis Ouglis Marque-pages Ouglis Marque-pages Ouglis Marque-pages Ougli		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A 1VR9A 1WR4A 1VR9A 1WMHA 1WPBG 1WMIA 1X6JA 1XB9A 1XE0B	101 177 255 235 666 175 255 149 103 121 83 168 88 88	$\begin{array}{c} 0.58\\ 0.29\\ 0.42\\ 0.41\\ 0.48\\ 0.30\\ 0.34\\ 0.28\\ 0.47\\ 0.42\\ 0.27\\ 0.42\\ 0.27\\ 0.37\\ 0.41\\ 0.43\\ 0.46\\ 0.61\end{array}$	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59 0.45 0.28 0.35 0.42 0.36 0.51 0.62	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82 1.00 0.79 0.82 0.82 0.86 0.81 0.76 0.81 0.90	Ficher Edition Affichage Allerà Marque-pages Ouglis Alde Image: Allerà Marque-pages Ouglis Allerà Image:		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A 1VR9A 1WR4A 1WPBG 1WMIA 1X6JA 1XB9A 1XE0B Mean	101 177 255 235 666 175 255 149 103 121 83 168 88 88 88 108	$\begin{array}{c} 0.58\\ 0.29\\ 0.42\\ 0.41\\ 0.48\\ 0.30\\ 0.34\\ 0.28\\ 0.47\\ 0.42\\ 0.27\\ 0.37\\ 0.41\\ 0.43\\ 0.46\\ 0.61\\ 0.42 \end{array}$	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59 0.45 0.28 0.35 0.42 0.36 0.51 0.62 0.44	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82 1.00 0.79 0.82 0.86 0.81 0.76 0.81 0.90 0.88	Ficher Edition Affichage Aller & Marque-pages Ouglis Aide Image: Aller & Marque-pages Ouglis Alde Image: Aller & Marque-pages Ouglis Aller Image: Aller & Marque-pages Ouglis Aller <tr< td=""></tr<>		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A 1VR9A 1WR4A 1WR9A 1WMHA 1WPBG 1WMIA 1X6JA 1XB9A 1XE0B Mean SD	101 177 255 235 666 175 255 149 103 121 83 168 88 88 88 108	$\begin{array}{c} 0.58\\ 0.29\\ 0.42\\ 0.41\\ 0.48\\ 0.30\\ 0.34\\ 0.28\\ 0.47\\ 0.42\\ 0.27\\ 0.42\\ 0.27\\ 0.37\\ 0.41\\ 0.43\\ 0.46\\ 0.61\end{array}$	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59 0.45 0.28 0.35 0.42 0.36 0.51 0.62	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82 1.00 0.79 0.82 0.82 0.86 0.81 0.76 0.81 0.90	Ficher Edition Michage Aller & Marque-pages Ouglis Adde Image: State of the problem of the		
1PXZA 1RKIA 1S7LA 1T70A 1TXOA 1V0ED 1V7BA 1VB5B 1VKCA 1VR4A 1VR9A 1WR4A 1VR9A 1WMHA 1WPBG 1WMIA 1X6JA 1XB9A 1XE0B	101 177 255 235 666 175 255 149 103 121 83 168 88 88 88 108	$\begin{array}{c} 0.58\\ 0.29\\ 0.42\\ 0.41\\ 0.48\\ 0.30\\ 0.34\\ 0.28\\ 0.47\\ 0.42\\ 0.27\\ 0.37\\ 0.41\\ 0.43\\ 0.46\\ 0.61\\ 0.42 \end{array}$	0.44 0.36 0.50 0.38 0.45 0.41 0.42 0.33 0.59 0.45 0.28 0.35 0.42 0.36 0.51 0.62 0.44	0.88 0.86 0.95 0.96 0.89 0.87 0.84 0.82 1.00 0.79 0.82 0.86 0.81 0.76 0.81 0.90 0.88	Ficher Edition Affichage Aller à Marque-pages Ouglis Aide Image: State of the fibro		

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.



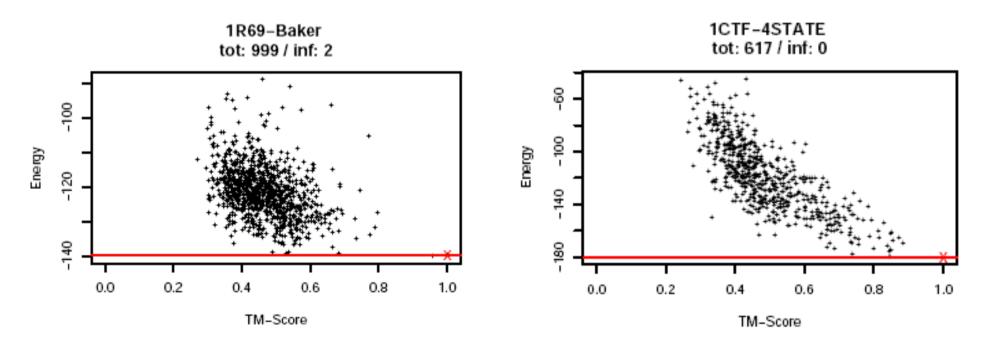
The greedy algorithm: the protein model is built linearly by overlapping optimal series of fragments (see above). At each step of the model reconstruction, the greedy algorithm builds all possible combinations of fragments, rank them according to an objective function, and keep only the **h** best solutions for the next iteration (**h** is called the heap).

Fragments overlap:



The greedy procedure:

Force field optimization [6]: to improve the capability of the force field to recognize native folds, we have optimized OPEP parameters on available decoys sets and our own ones. Here are presented an Energy VS TM-Score [7] plot on 2 validation decoys sets.

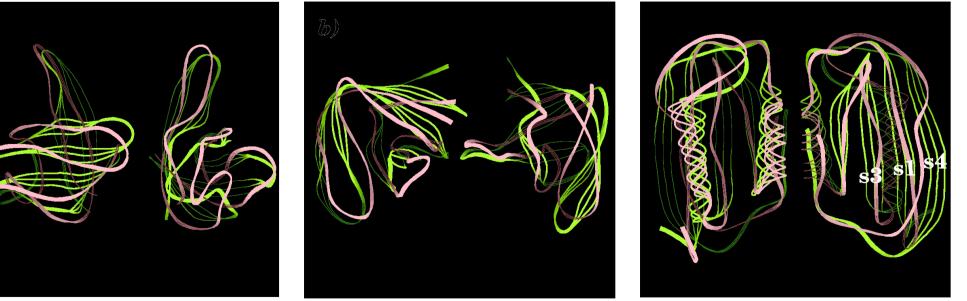


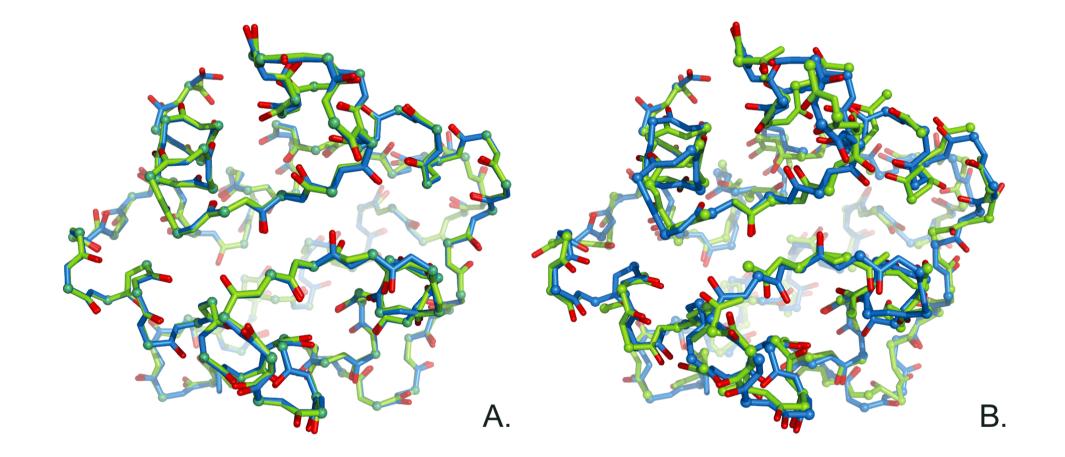
Greedy algorithm improvements: the basic algorithm has been improved on 3 points [2]:

(i) the algorithm is now *stochastic* : a part of the heap is now randomly chosen (a local minimum is not a global one). (ii) we add *pre-filters* (rebuilding of small parts of the protein to avoid some transitions).

(iii) the algorithm is *iterated* (from N to C terminus and then from C to N terminus).

Here are presented 3 predicted models (Go criterion) superposed to the experimental structure (*in green*).

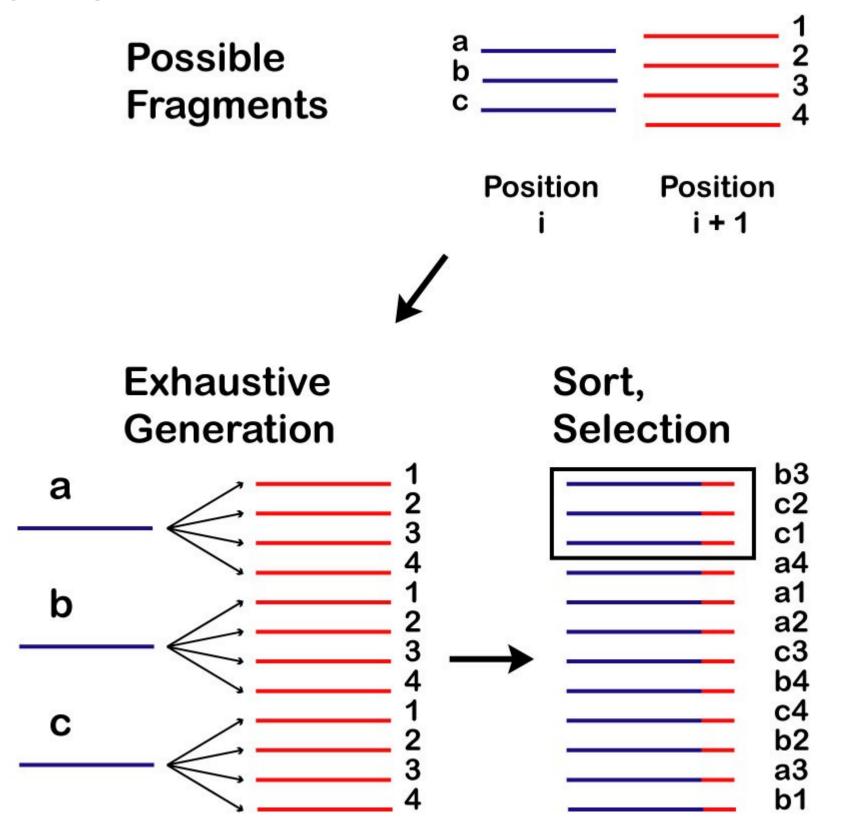




SABBAC rebuilding example, 10MD. (A) Native and SABBAC rebuilt structures. (B) Native and SABBAC rebuilt structures from alpha-carbon trace perturbed by 0.8A on average. The native structure is represented in blue.

Perspectives: two different directions are now considered : ab initio model generation, and comparative modeling.

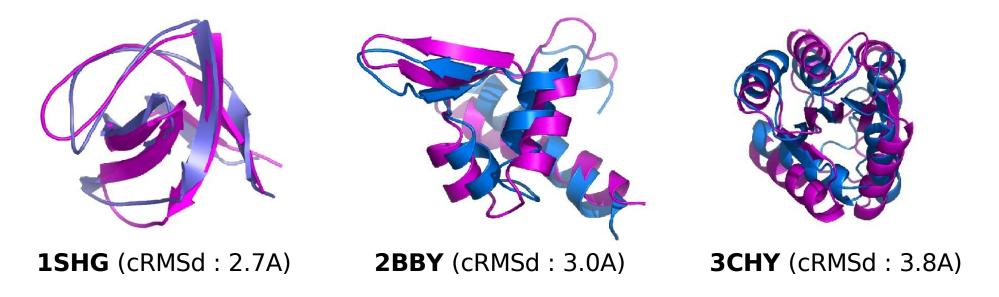
For the comparative modeling case, given an alignment, we are able to fit some parts of the template to generate a model. A new term has been added to the energetic function, minimizing the deviation from the template. The method is still in development.



1CSP (cRMSd : 3.1A)

1SHG (cRMSd : 3.2A) **2ACY** (cRMSd : 3.4A)

Using only secondary structure information, and considering all the possible letters for non structured regions (i.e. loops), it is possible to reduce search complexity by considering the dependence between consecutive local conformations deduced from the HMM [**3**]. The procedure can reproduce 20 protein structures of 50-164 amino acids within 2.7 to 6.5A cRMSd. Three examples are presented here (experimental structures are in magenta):



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] Tuffery P, Guyon F, Derreumaux P. Improved greedy algorithm for protein structure reconstruction. J Comput Chem. 2005 Apr 15;26(5):50613.

[3] 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.

[4] Santini S, Wei G, Mousseau N, and Derreumaux P. Exploring the Folding Pathways of Proteins through Energy Landscape Sampling: Application to Alzheimer's Amyloid Peptide. Internet Electron. J. Mol. Des. 2003, 2, 564577

[5] 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.

[6] Maupetit J, Tuffery P and Derreumaux P. A refined knowledge-based force field for protein folding and structure prediction. In preparation.

[7] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins.* 2004 Dec 1;57(4):702-10.