A Greedy Algorithm for Protein Structure Reconstruction: *Improvements and Applications*

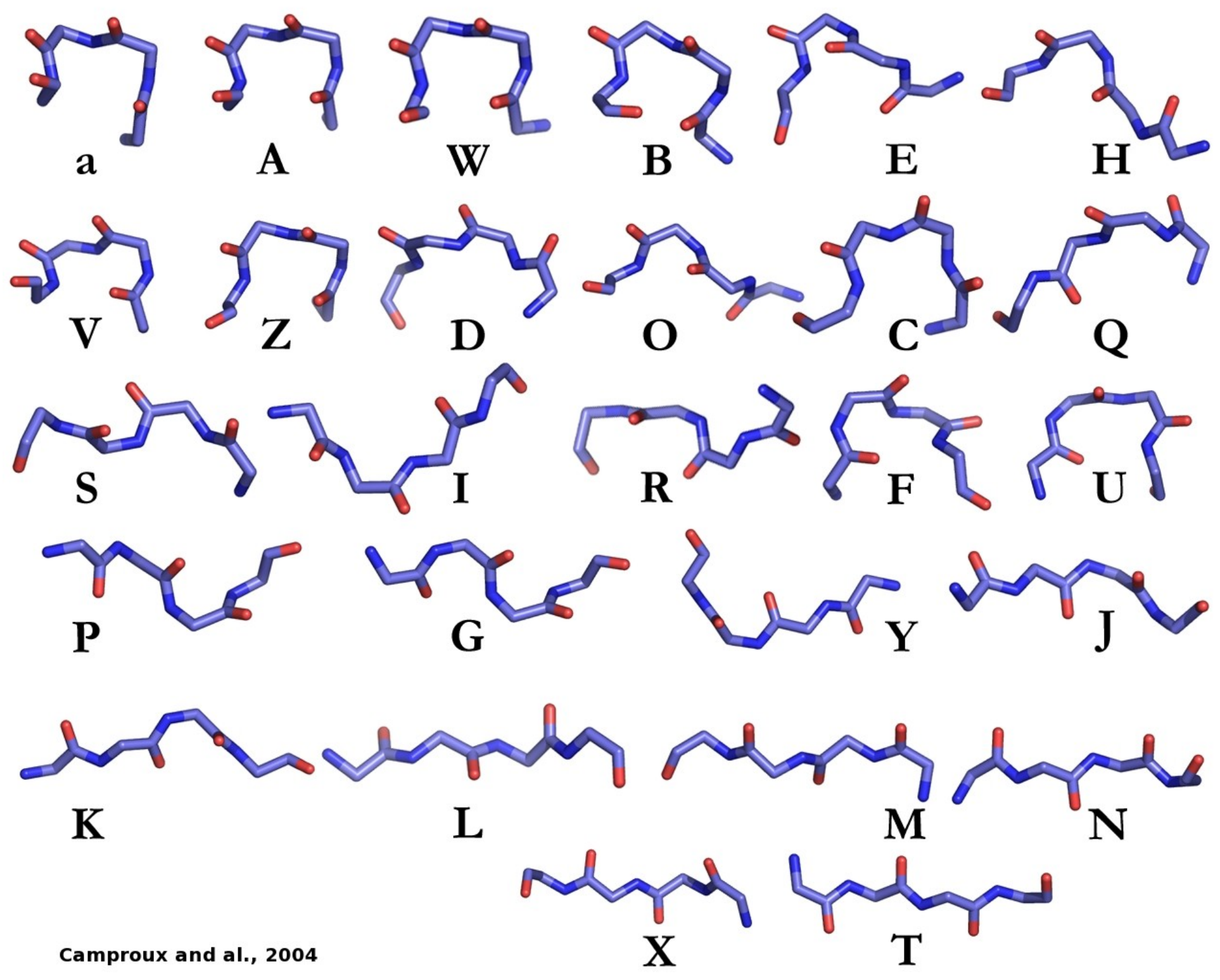
Julien Maupetit[1], **Philippe Derreumaux**[2] **and Pierre Tufféry**[3]

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

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

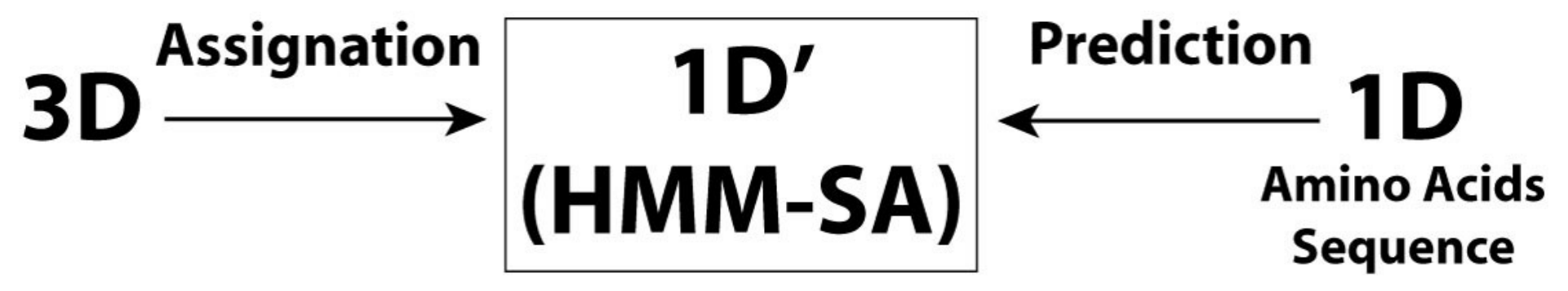
Goal: Large scale genome sequencing projects bring sequence databases to grow exponentially. Even if the number of known proteins structures increases rapidly, only few of the detected ORFs correspond to proteins having an experimentally resolved 3D structure. For such reason, one needs to develop high 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.

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).



Camproux and al., 2004

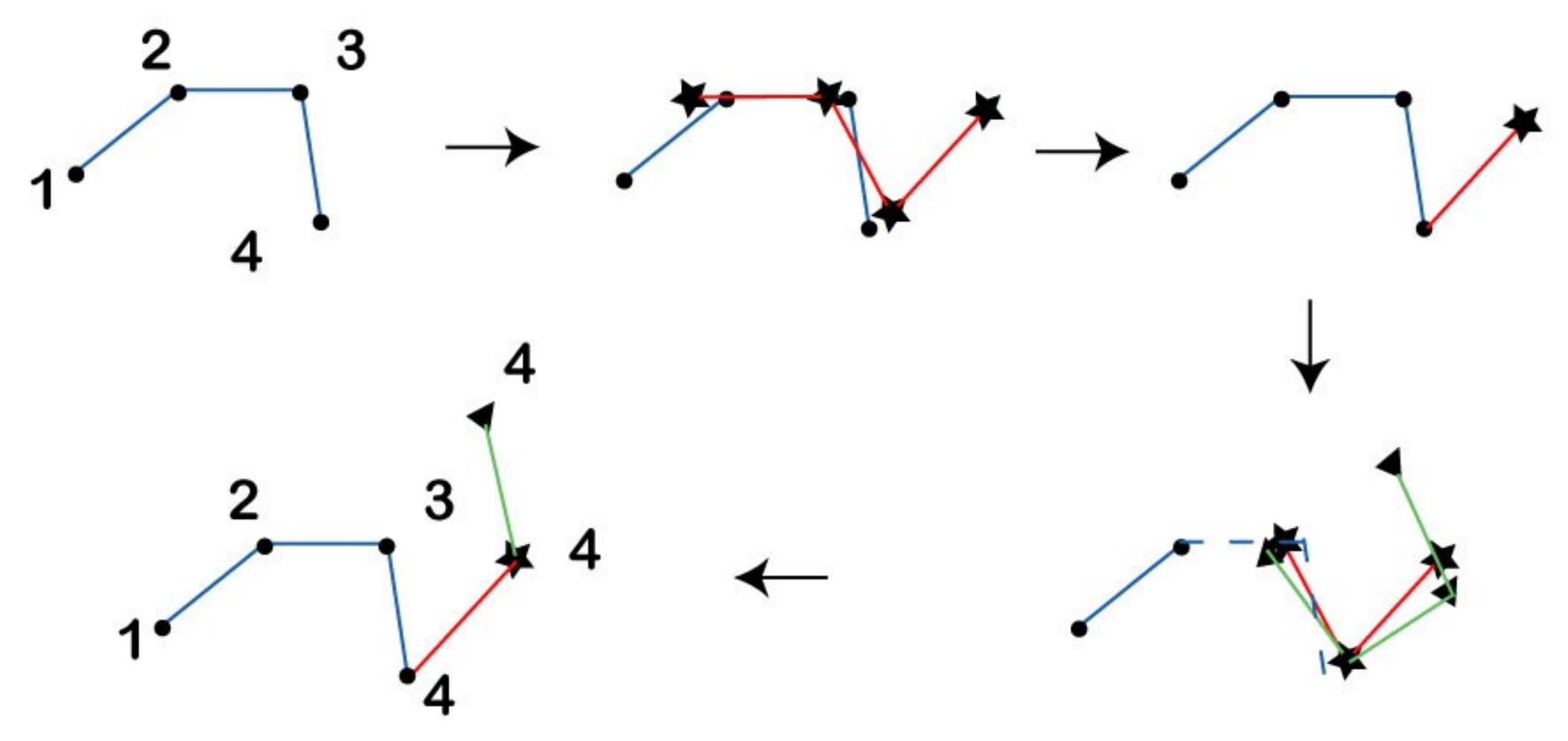
Generate HMM-SA description of the model: there are two different ways to determine the collection of fragments to use at each position of the protein : *prediction* or *assignation*. Prediction starts from the amino acid sequence to propose structural alphabet letters. Assignation starts from known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



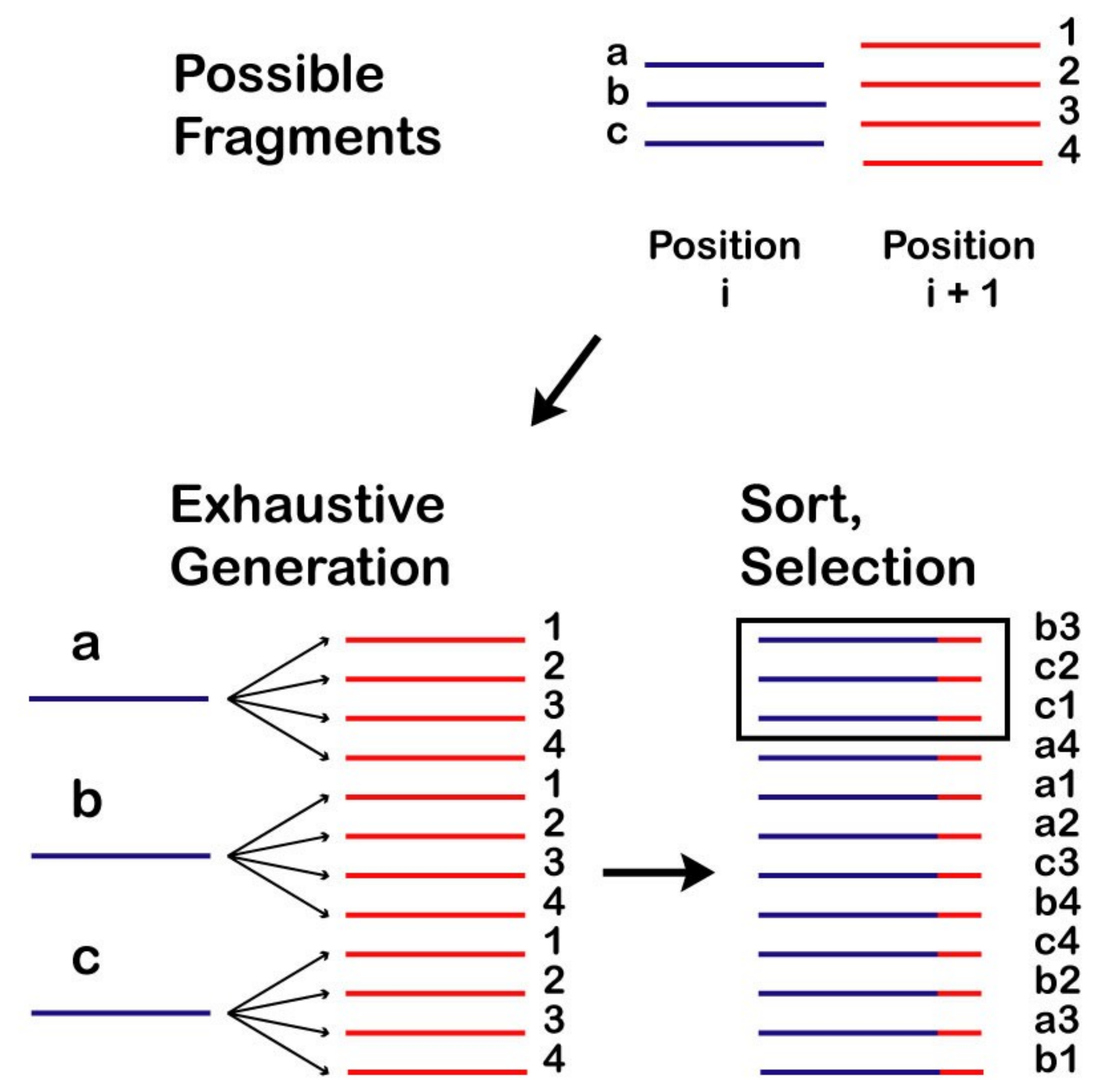
3D
Assignation
1D'
(HMM-SA)
Prediction
1D
Amino Acids
Sequence

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:



Possible
Fragments

a
b
c
1
2
3
4

Position
i

Position
i + 1

Exhaustive
Generation

Sort,
Selection

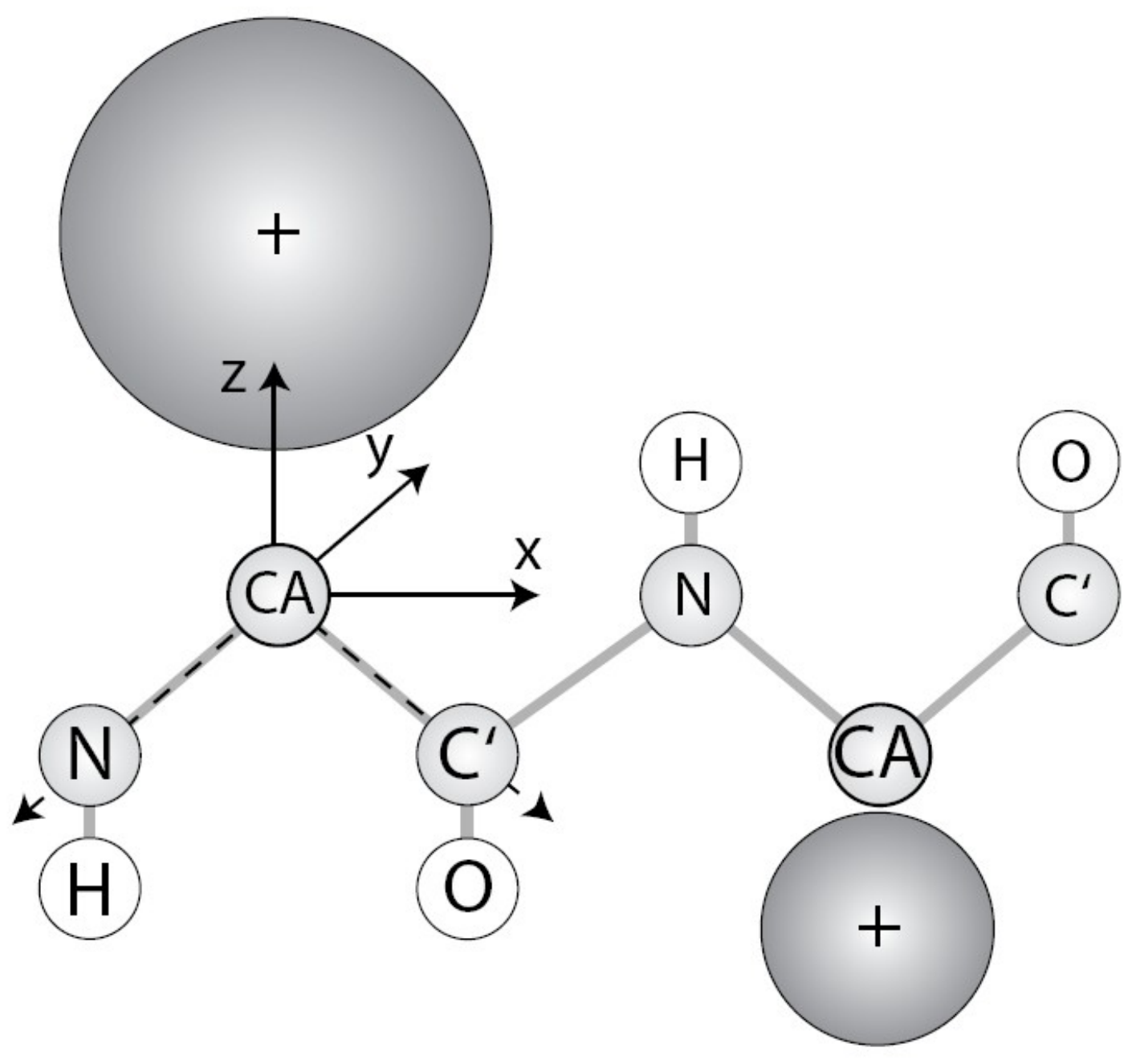
Objective function: to test the capability of the algorithm to reconstruct protein structures, alpha carbons *RMSd* (cRMSd) criterion was first used. It has been shown that greedy algorithm is able to rebuild protein structures at less than 1A cRMSd [**2**].

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**].

$$E_{\mathrm{Go}} = \sum_{j>i} \Delta_{ij} B_{ij} + E_{\mathrm{rep}}$$

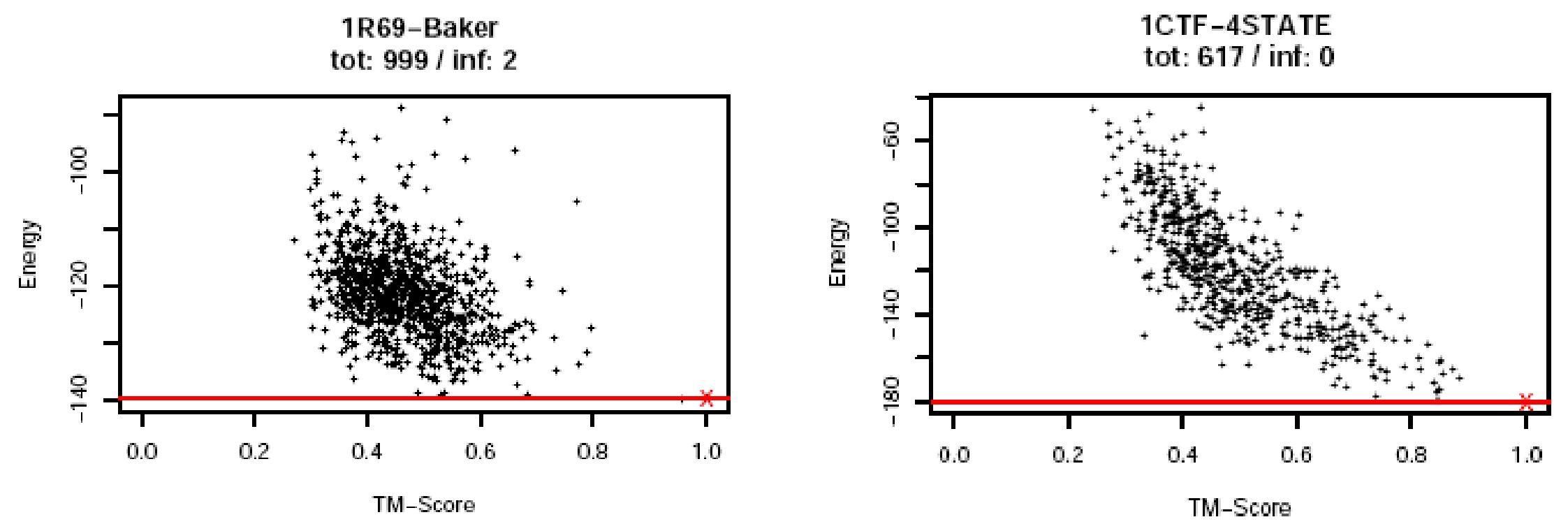
Finally, in an *ab initio* context, we started to implement a simplified version of the coarse grained *OPEP* [**4**] force field (*sOPEP*) with the following energetic terms:

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



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.

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.



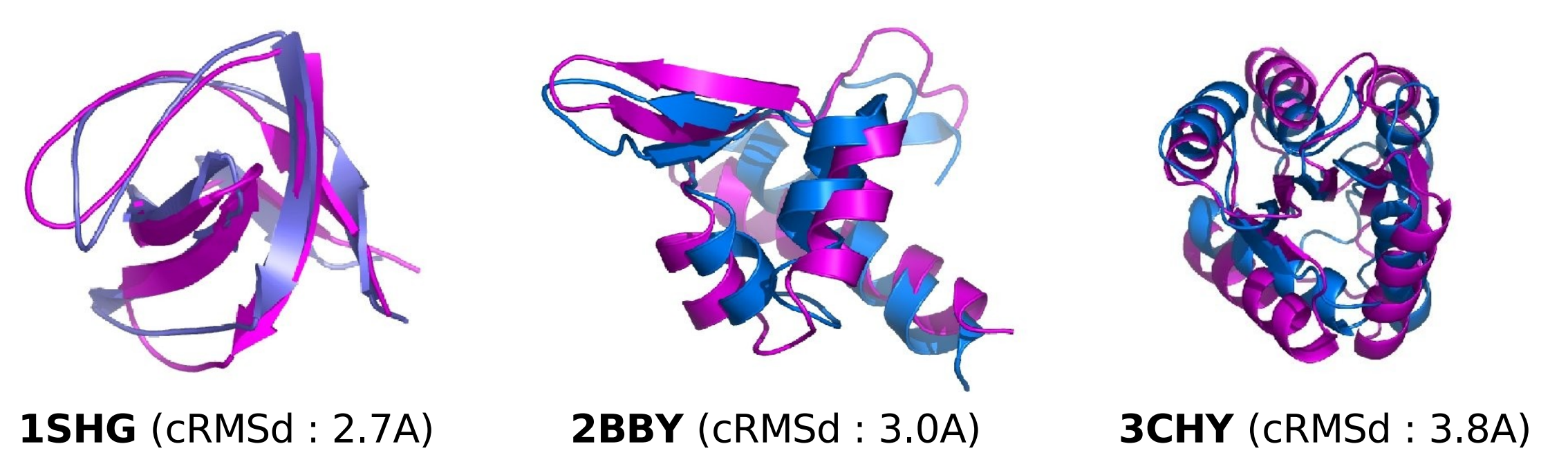
Greedy algorithm improvements: the basic algorithm has been improved on 3 points [**2**]:

(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*).

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):

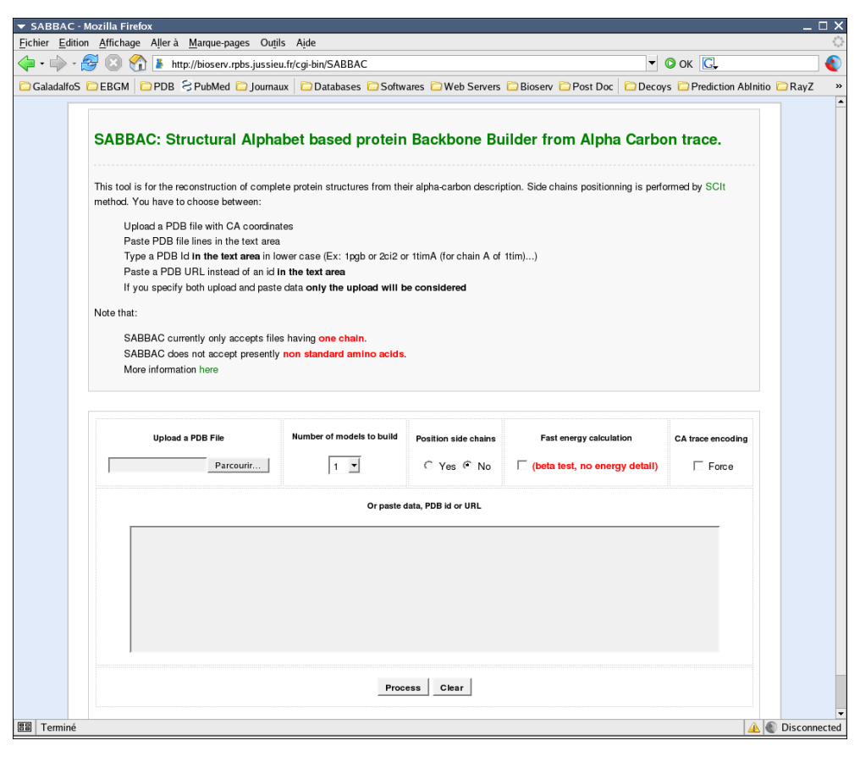


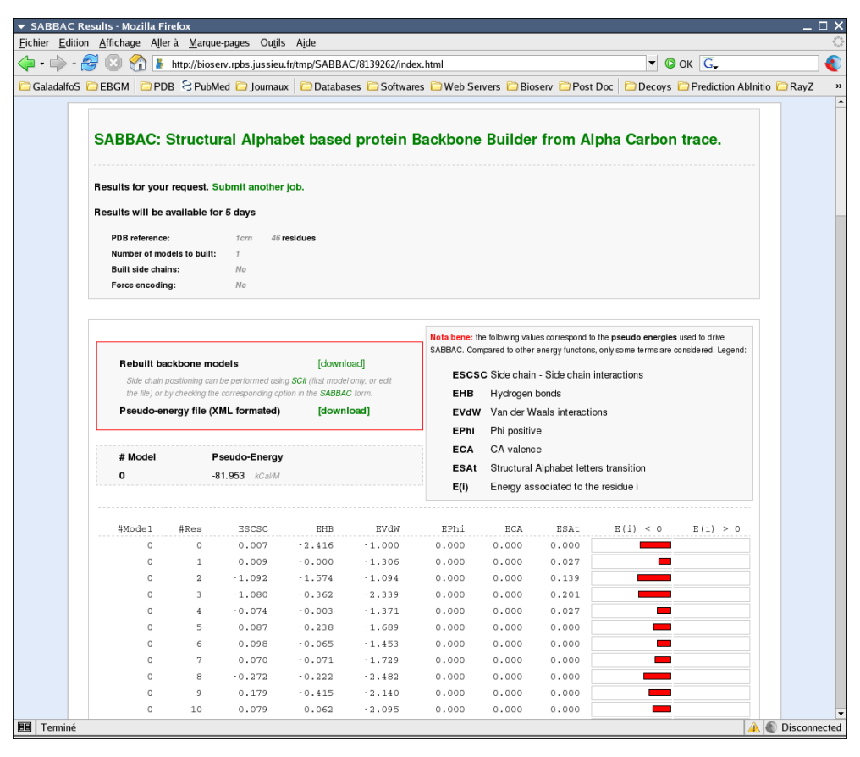
1SHG (cRMSd : 2.7A) **2BBY** (cRMSd : 3.0A) **3CHY** (cRMSd : 3.8A)

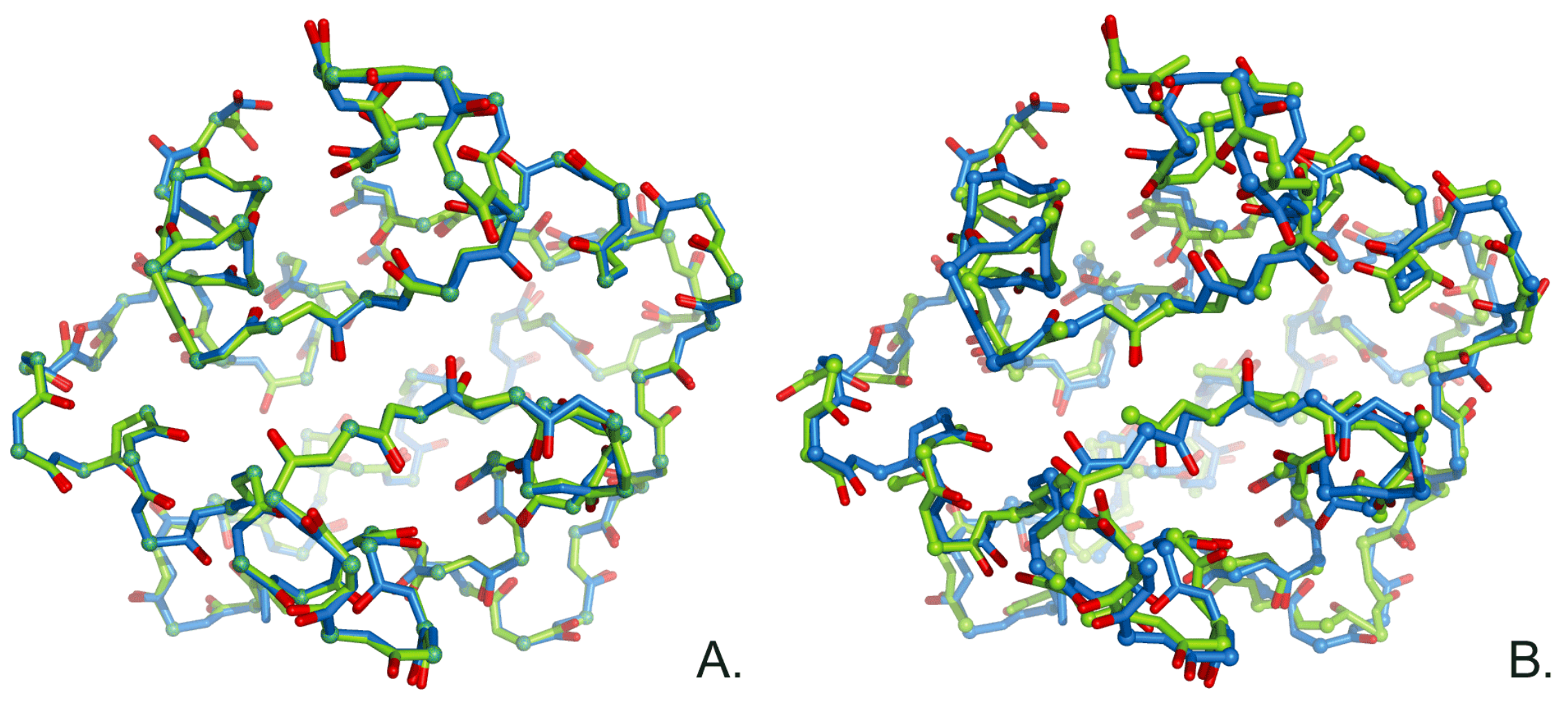
SABBAC (**S**tructural **A**lphabet based **B**ack**B**one reconstruction from **A**lpha-**C**arbon trace) [**5**]: based on our greedy algorithm, we implemented a web tool to rebuild protein backbone from alpha-carbon trace. Alpha-carbon coordinates remain unaffected. SABBAC simply positions the missing 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

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

PDB	*n* Residues	Main chain RMSd SABBAC	MaxSprout	*bb*
Adcock subset				
111M	154	0.29	0.42	0.91
1CTF	68	0.43	0.73	0.85
1IGD	61	0.36	0.44	0.68
1OMD	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	0.45	0.86
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
Mean		0.41	0.47	0.89
SD		0.09	0.10	0.08
PDB newcomers subset				
1PXZA	346	0.55	0.54	0.96
1RKIA	101	0.58	0.44	0.88
1S7LA	177	0.29	0.36	0.86
1T70A	255	0.42	0.50	0.95
1TXOA	235	0.41	0.38	0.96
1V0ED	666	0.48	0.45	0.89
1V7BA	175	0.30	0.41	0.87
1VB5B	255	0.34	0.42	0.84
1VKCA	149	0.28	0.33	0.82
1VR4A	103	0.47	0.59	1.00
1VR9A	121	0.42	0.45	0.79
1WMHA	83	0.27	0.28	0.82
1WPBG	168	0.37	0.35	0.86
1WMIA	88	0.41	0.42	0.81
1X6JA	88	0.43	0.36	0.76
1XB9A	108	0.46	0.51	0.81
1XE0B	107	0.61	0.62	0.90
Mean		0.42	0.44	0.88
SD		0.10	0.09	0.07
GLOBAL				
Mean		0.41	0.45	0.87
SD		0.09	0.10	0.07







SABBAC rebuilding example, 1OMD. (**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.

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.