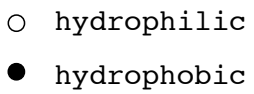
Proteus
Exploration de l' espace séquence-rotamère
Comparaison d'algorithmes

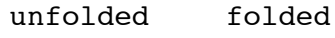
David Mignon 11/05/15

Biomolecular simulation and design

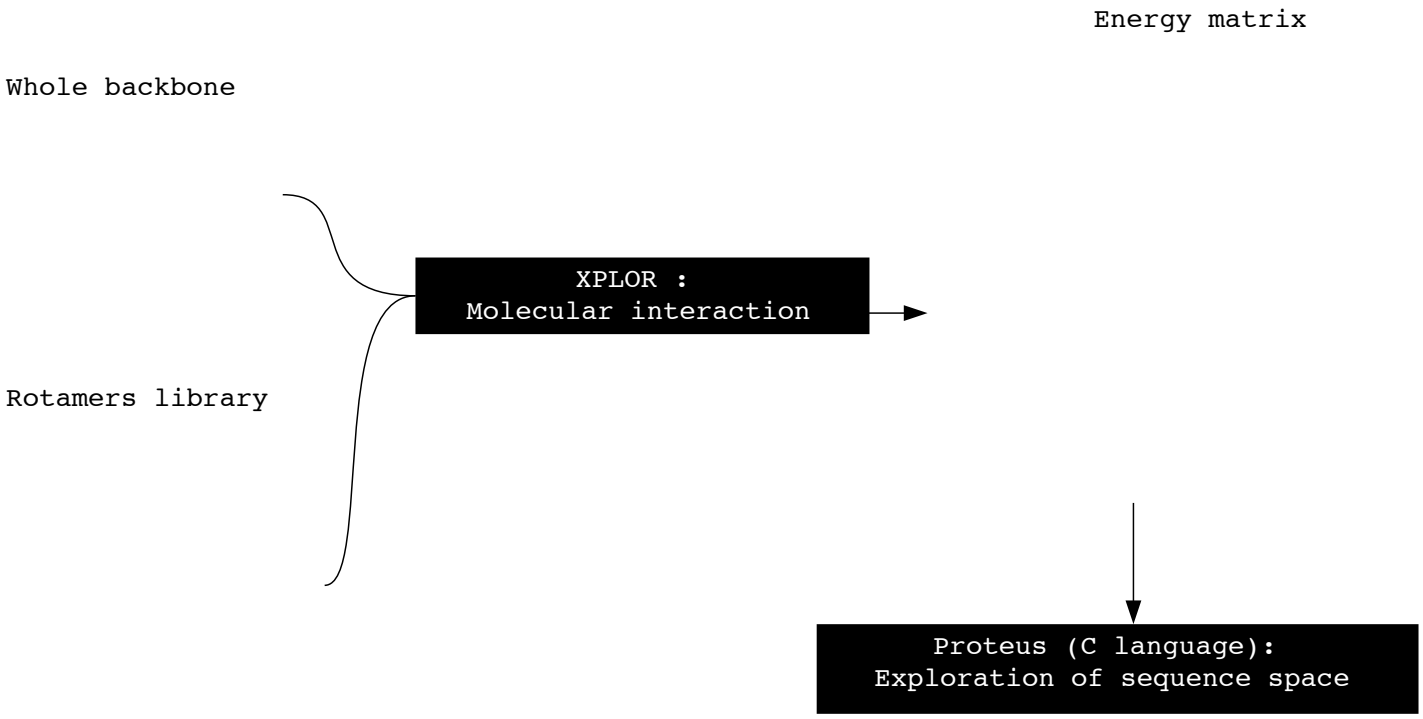
- Conformational space
  - Particular structure for backbone
  - Side chains rotamers
  - Simple model of unfolded state (important for stability)



- Energy function : pairwise decomposition
  - Intraprotein = $E_{bb} + \Sigma E_{ii} + \Sigma E_{ij}$
  - Solvant = dielectric model



Energy matrix or lookup table



Comparaison de méthodes d'exploration des séquences

A. Présentation des algorithmes

- Algorithme de recherche exacte: Toulbar2
- Une heuristique spécifique à la structure de l'espace
- méthodes probabilistes

1. Monte Carlo
2. Replica Exchange

B. Ensemble des tests

- méthodes
- résultats
- analyses

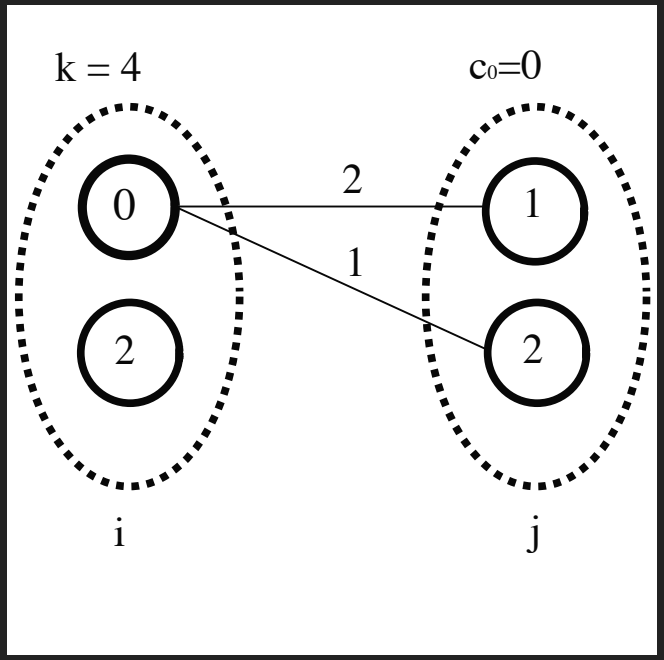
Une méthode exacte par optimisation combinatoire

La décomposition par paire de notre fonction d'énergie permet une représentation des énergies sous forme d'

"un réseau de fonctions de coûts".

interaction entre acides aminés <=> une arête dans le réseau

une énergie d'un rotamère <=> un nœud dans le réseau



Le logiciel toulbar2 (D. Allouche,S. de Givry, Schiex)

1. Un ensemble de transformations qui préservent l'équivalence des problèmes est appliqué sur le réseau.
2. Un minorant m et un majorant M du minimum global d'énergie sont mis à jour après transformation.
3. Un arbre d'états est "élagué" à partir de m et M.
4. retour en 1.

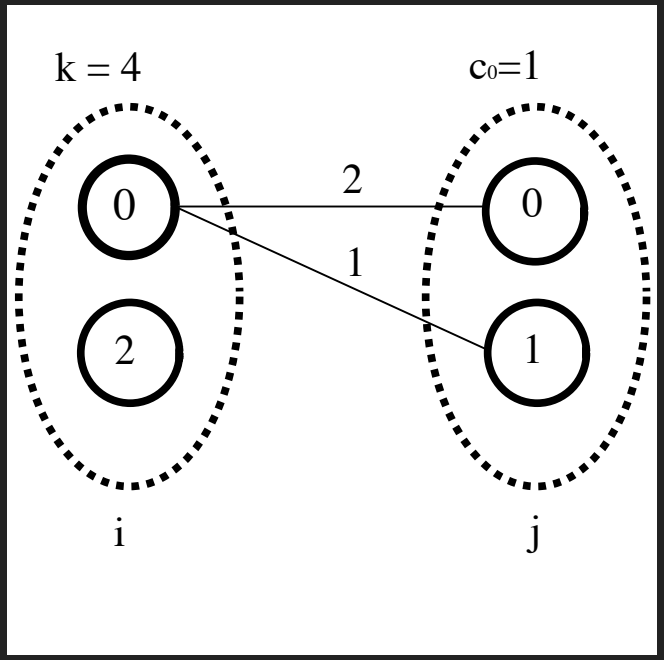
Une méthode exacte par optimisation combinatoire

La décomposition par paire de notre fonction d'énergie permet une représentation des énergies sous forme d'

"un réseau de fonctions de coûts".

interaction entre acides aminés <=> une arête dans le réseau

une énergie d'un rotamère <=> un nœud dans le réseau



Le logiciel toulbar2 (D. Allouche,S. de Givry, Schiex)

1. Un ensemble de transformations qui préservent l'équivalence des problèmes est appliqué sur le réseau.
2. Un minorant m et un majorant M de minimum global d'énergie sont mis à jour après transformation.
3. Un arbre d'états est "élagué" à partir de m et M.

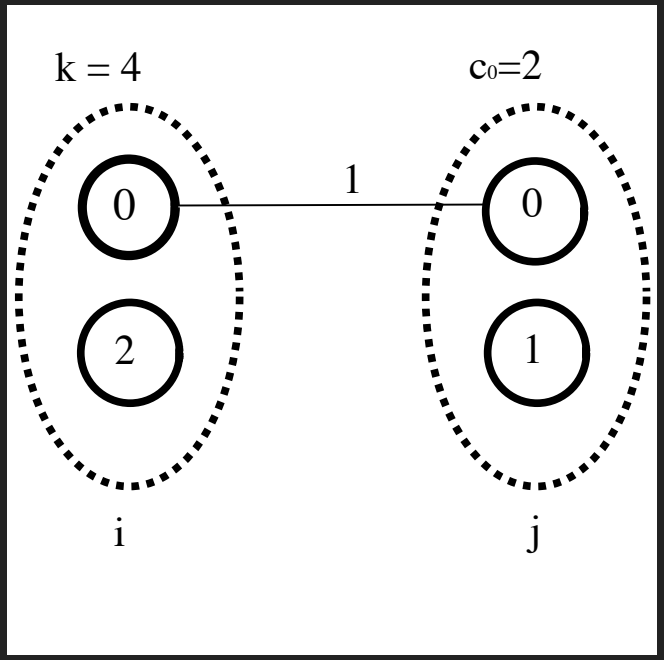
Une méthode exacte par optimisation combinatoire

La décomposition par paire de notre fonction d'énergie permet une représentation des énergies sous forme d'

"un réseau de fonctions de coûts".

interaction entre acides aminés <=> une arête dans le réseau

une énergie d'un rotamère <=> un nœud dans le réseau



Le logiciel toulbar2 (D. Allouche,S. de Givry, Schiex)

1. Un ensemble de transformations qui préservent l'équivalence des problèmes est appliqué sur le réseau.
2. Un minorant m et un majorant M de minimum global d'énergie sont mis à jour après transformation.
3. Un arbre d'états est "élagué" à partir de m et M.

Une méthode heuristique spéficique à la structure de l'espace (Wernisch,Wodak)

Le but est la d'obtenir un ensemble de séquences-rotamères qui approchent le minimum globale.

L'algorithme effectue une optimisation à chaque position:

```
1  Pour chaque cycle heuristique
2    Une séquence-rotamère S est choisie aléatoirement
3    Tant que l'énergie de S est améliorée
4          Pour i allant de la première position de S jusqu'à la dernière
5                S est fixée sauf à la position i
6                Le meilleur rotamère possible en i est determiné
7                Ce rotamère est utilisé pour fixer S en i
8          fin de Pour
9    fin de Tant que
10   S est sauvegardée
11 fin de Pour
```

Le Monte Carlo

algorithme Metropolis-Hastings

Le but est de générer une collection d'état échantillonné selon la distribution de Boltzmann.

$$p(état) \propto e^{(\frac{-E}{RT})}$$

L'algorithme définit une chaîne de Markov pour laquelle:

- la distribution de probabilité des états est stationnaire.

  C'est garantie par la balance détaillée.

- Il n'y a qu'une seule distribution stationnaire.

  C'est garantie par le caractère ergodique de la chaine.

Le Monte Carlo

algorithme Metropolis-Hastings

1 Une séquence-rotamère S_0 est choisie aléatoirement
2 Pour i allant du premier pas de la trajectoire jusqu'au dernier
3 A partir d'une probabilité conditionnelle select(. , S), une proposition S'_i est choisie par un tirage qui suit select(.,S_i)
4 on calcule la probabilité d'acceptation: acc=exp(βΔE)select(S_i,S'_i)/select(S'_i,S_i)
5 si acc ≥ 1 alors $S_{i+1}=S'_i$
sinon alors $S_{i+1} = S'_i$, avec la probabilité acc
sinon $S_{i+1} = S_i$
6 fin de Pour

Dans proteus select(S',S) est une combinaison de changements de rotamères et/ou de mutations.

C'est une fonction symétrique!

Le Monte Carlo

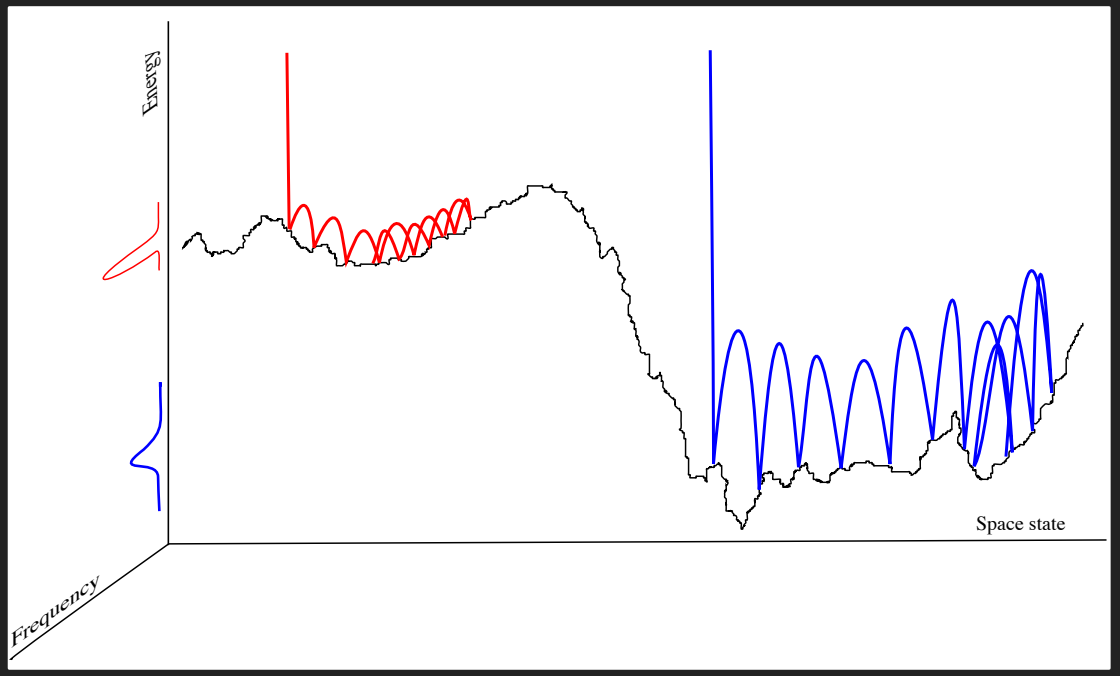
algorithme Metropolis-Hastings

Avantages:

- ensemble
- propriétés physique de la trajectoire

inconvénients:

- Non exacte / stochastique
- exploration peut être inefficace



Repliqua Exchange

Objectif: améliorer le Monte Carlo pour

pouvoir franchir plus souvent les barrières énergétiques tout en conversant les qualités de l'échantillon.

1 Lancement en parallèle de N marcheurs Monte Carlo aux températures ordonnées (t_1,..,t_N)

2 Tous les P pas

3 i est choisie aléatoirement entre 1 et N
ce qui sélectionne les marcheurs aux températures t_i et t_{i+1}.

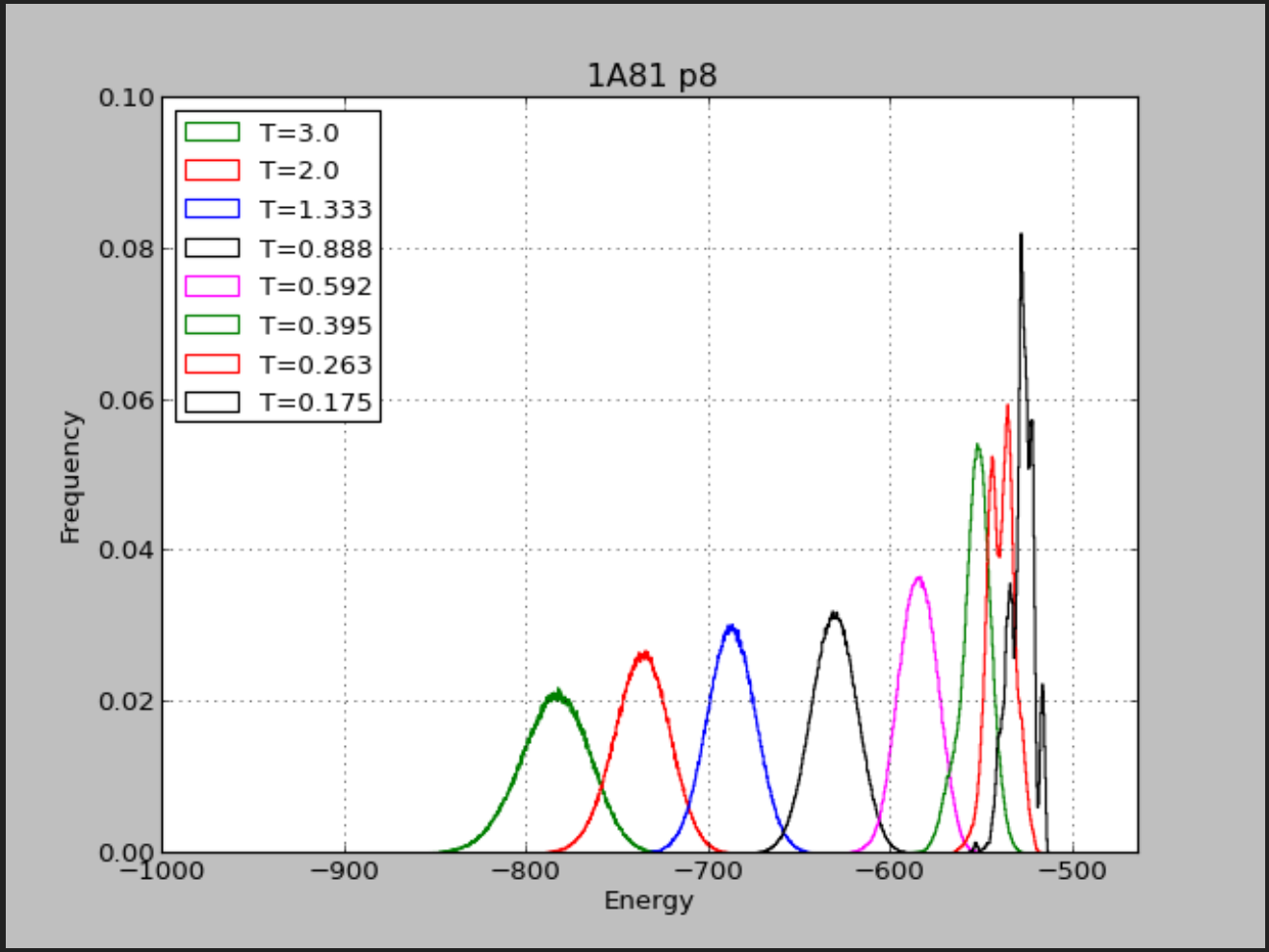
4 La probabilité d'acceptation suivante est calculée
$acc=\exp((E_i-E_{i+1})(1/t_i-1/t_{i+1}))$

5 si $acc \geq 1$ alors les températures sont échangées
sinon alors les températures sont échangées , avec la probabilité acc

6 fin de la trajectoire

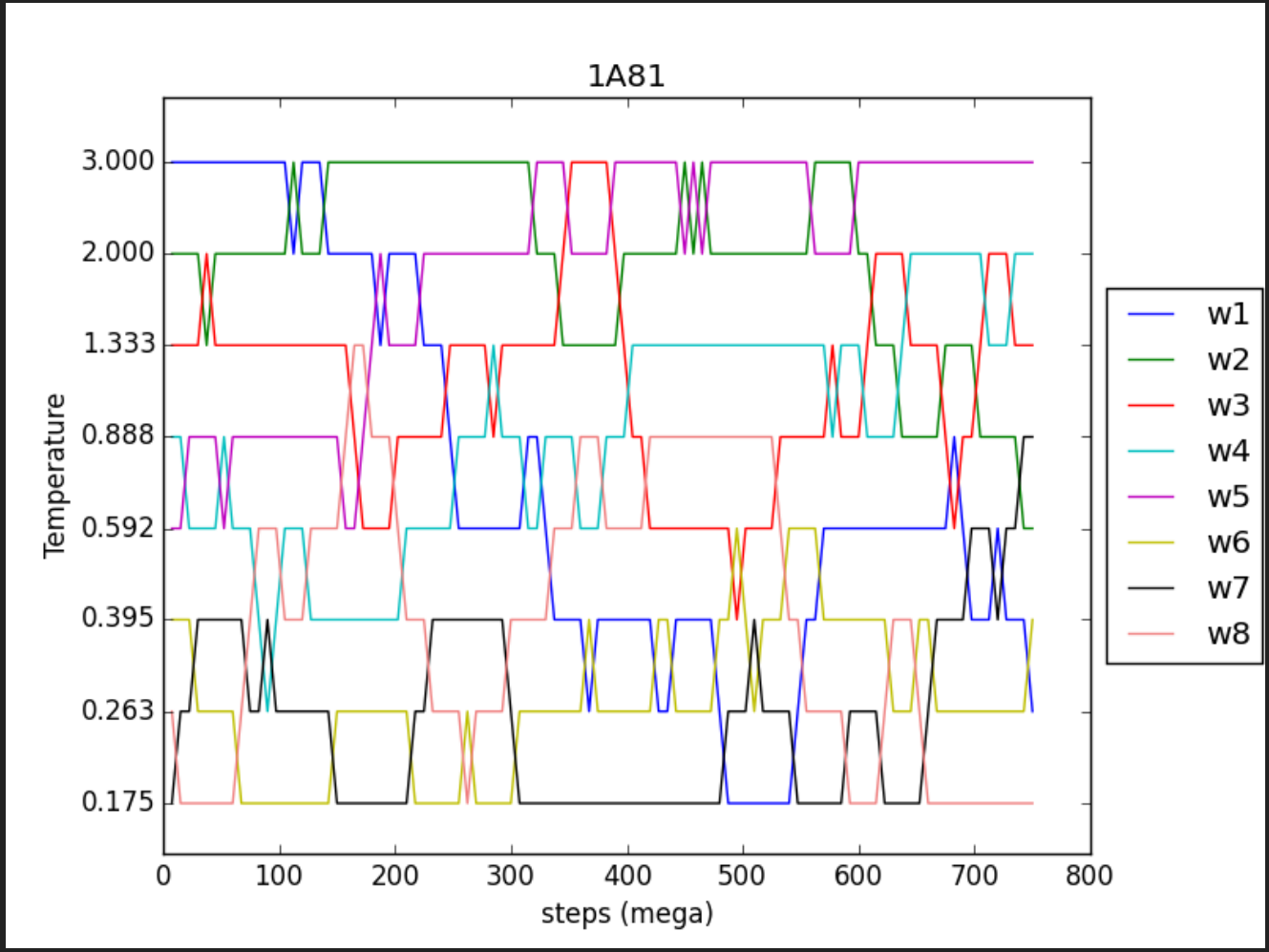
Repliqua Exchange

Comportement de l'algorithme



Repliqua Exchange

Comportement de l'algorithme



Ensemble de tests

1. 8 protéines de 57 jusqu'à 109 résidus:

   1A81(58),1ABO(98),1BM2(57),1CKA(57),1G90(91), 1M61(109),104C(104),1R6J(82),2BYG(97)

2. Méthodes

   - Toulbar2: temps d'exécution max 24h, en cas d'échec relance avec une deuxième configuration

   - Heuristique: 110 000 cycles (environ 24h de calculs pour le plus long test)

   - Monte Carlo: 6 milliards de pas (environ 24h de calculs pour le plus long test)

      . à chaque pas, 2 modifications de rotamères et 1 mutation en moyenne tous les 10 pas

      . température: 0.2

   - Replica Exchange: 6 milliards de pas cumulés sur tous les marcheurs

      . 4 marcheurs -> températures 10<->0.01 ou 2<->0.25

      . 8 marcheurs -> températures:10<->0.003 ou 3<->0.175, swap tests: 100 ou 750,...

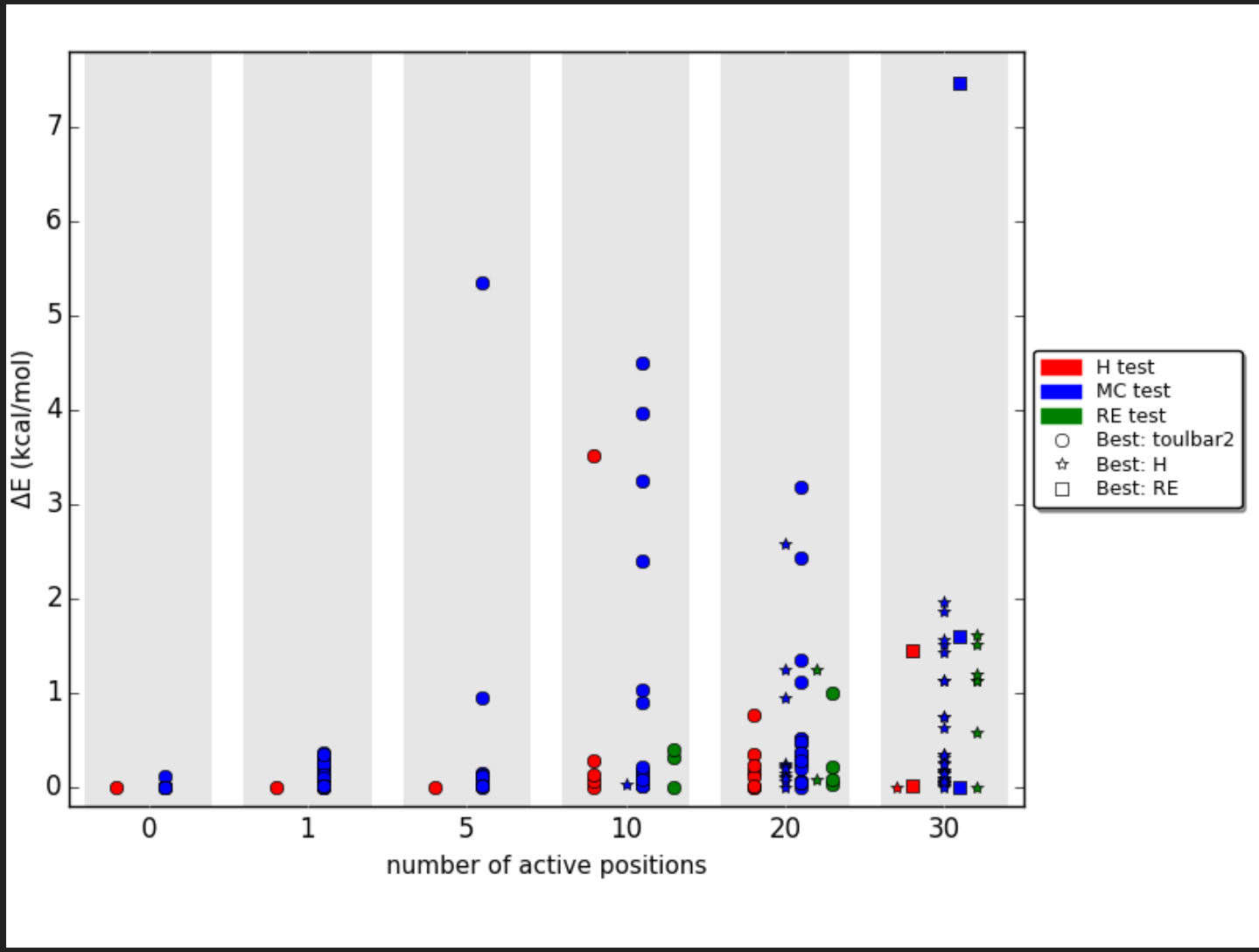
3. types de tests: séquences fixés,mutation 1 position,mutation multi-positions,tous actifs.

4. Analyses: atteinte du minimum global, étude au voisinage, structures

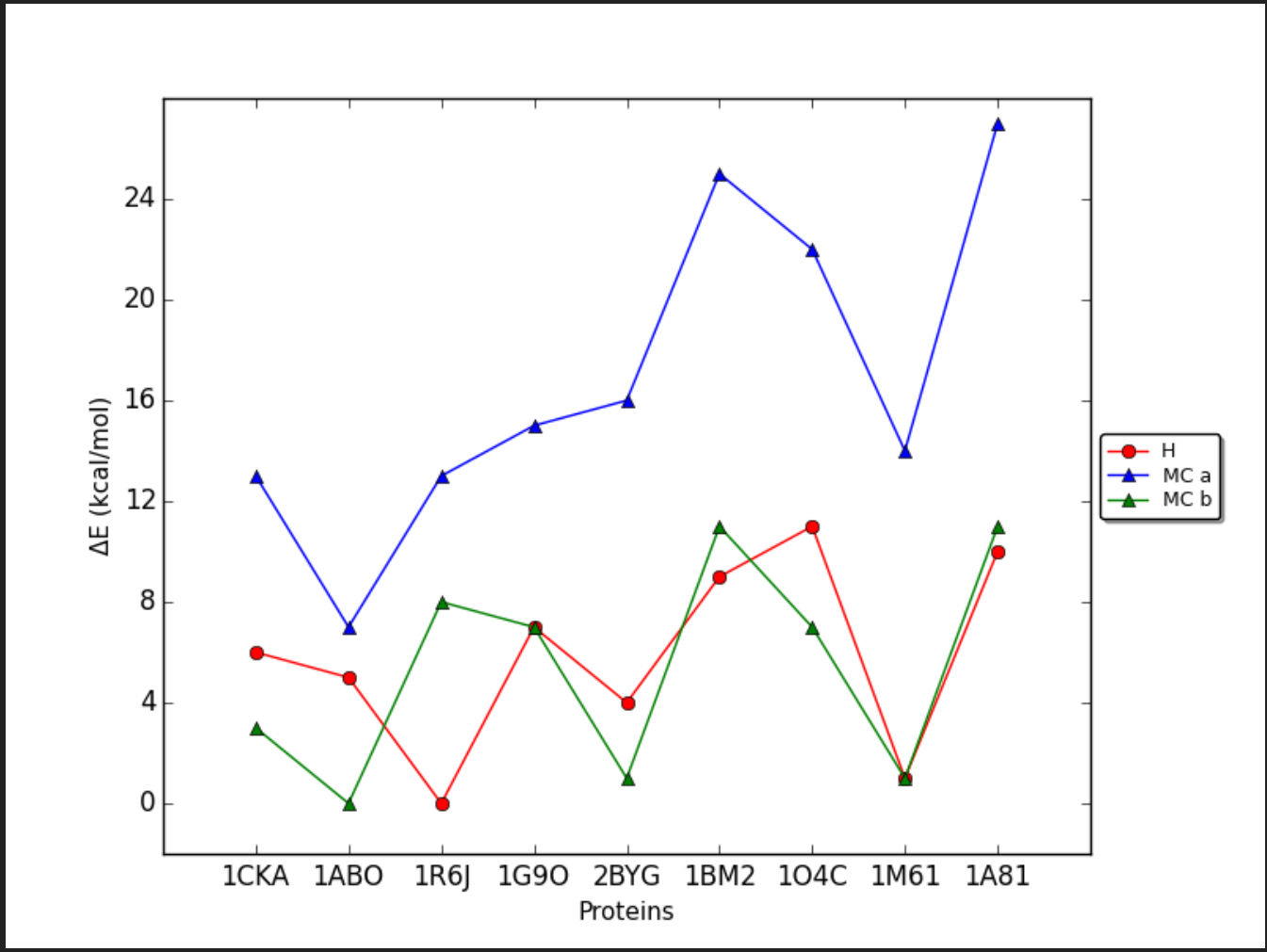
Résultats pour 10 et 20 positions actives

	10 positions actives				20 positions actives			
Protéine	GMEC	H	MC	RE	GMEC	H	MC	RE
1A81 1	-583.9354	0	0		-566.9106	0	-0.3275	
1A81 2	-581.7802	0	0		-564.6618	-0.1705	-2.4355	-1.0069
1A81 3	-587.4392	-0.0001	-0.1595		-572.9780	0	-0.4640	
1A81 4	-589.1322	0	-0.0317		-570.3480	-0.3568	-0.5128	
1A81 5	-578.2558	0	-0.0563		-571.2480	-0.7658	-0.5088	
1ABO 1	-309.1670	-0.0675	-0.9054		-299.6592	-0.1205	-1.1159	-0.2153
1ABO 2	-308.8387	0	0		no	-298.3854	0	
1ABO 3	-303.8520	0	0		no	-298.3854	0	
1ABO 4	-310.0087	0	-0.0128		no	-297.8545	-0.0076	
1ABO 5	-301.6727	0	0		no	-297.8009	-0.9483	
1BM2 1	-549.8638	0	-0.0950		-526.0936	0	-0.0619	
1BM2 2	-541.5944	0	0		no	-525.3588	-0.0725	
1BM2 3	-543.7434	0	0		-534.3860	-0.0230	-0.4763	
1BM2 4	-549.0453	0	0		no	-526.8307	-2.5883	-0.0789
1BM2 5	-544.1447	0	-0.1082		-535.3334	-0.2396	-0.3746	
1CKA 1	-305.8477	0	0		-295.6311	0	0	
1CKA 2	-309.9886	0	0		-295.8571	0	0	
1CKA 3	-304.6618	0	0		-293.8687	0	0	
1CKA 4	-302.4894	0	0		no	-293.8687	0	
1CKA 5	-299.2329	-0.2859	-3.2525	0	no	-293.4203	0	
1G9O 1	-466.6764	0	0		no	-451.4604	-1.2525	-1.2525
1G9O 2	-478.8797	0	0		no	-453.2355	-0.2487	
1G9O 3	-477.2503	-0.1366	0		no	-453.2474	-0.2177	
1G9O 4	-470.6458	0	0		no	-456.3751	-0.2275	
1G9O 5	-464.8659	0	-3.9599		no	-456.7331	-0.1455	
1M61 1	-550.0699	0	-0.0776		-528.0700	0	0	
1M61 2	-538.6026	-3.5105	-4.5062	0.3215	-528.7653	0	0	
1M61 3	-552.2673	0	0		-530.0684	0	0	
1M61 4	-550.0553	0	0		-534.5248	0	0	
1M61 5	-553.6559	0	-0.0432		-548.0096	0	-0.2521	
1O4C 1	-587.4665	0	-0.1121		no	-0.2775	-574.0737	
1O4C 2	-585.8545	0	-0.1046		no	-574.8584	-0.1963	
1O4C 3	-580.3505	0	-0.1519		-573.6314	0	-0.3461	
1O4C 4	-587.1548	0	-0.1545		-575.8667	0	-0.3640	
1O4C 5	-590.2650	0	-0.1753		no	-573.3479	-0.1141	
1R6J 1	-448.8351	0	-2.4022	-0.3986	-440.7417	0	-0.2604	
1R6J 2	-448.4631	0	-1.0398		-437.2537	0	-0.0071	
1R6J 3	-450.3950	0	-0.0106		-439.4335	0	-0.0537	
1R6J 4	-451.7211	0	0		-439.9135	0	-0.0537	
1R6J 5	-450.9943	0	-0.0162		-438.0222	0	-0.0735	
2BYG 1	no	-505.6397	-0.0337		-496.2991	0	-3.1878	-0.0257
2BYG 2	-504.7389	0	0		-494.8723	0	-0.0524	
2BYG 3	-504.3048	0	-0.0833		-494.8723	0	-1.3564	-0.0826
2BYG 4	-504.3466	0	-0.2149		-495.9213	0	-0.1968	
2BYG 5	-491.6095	0	0		no	-497.5123	-0.0933	

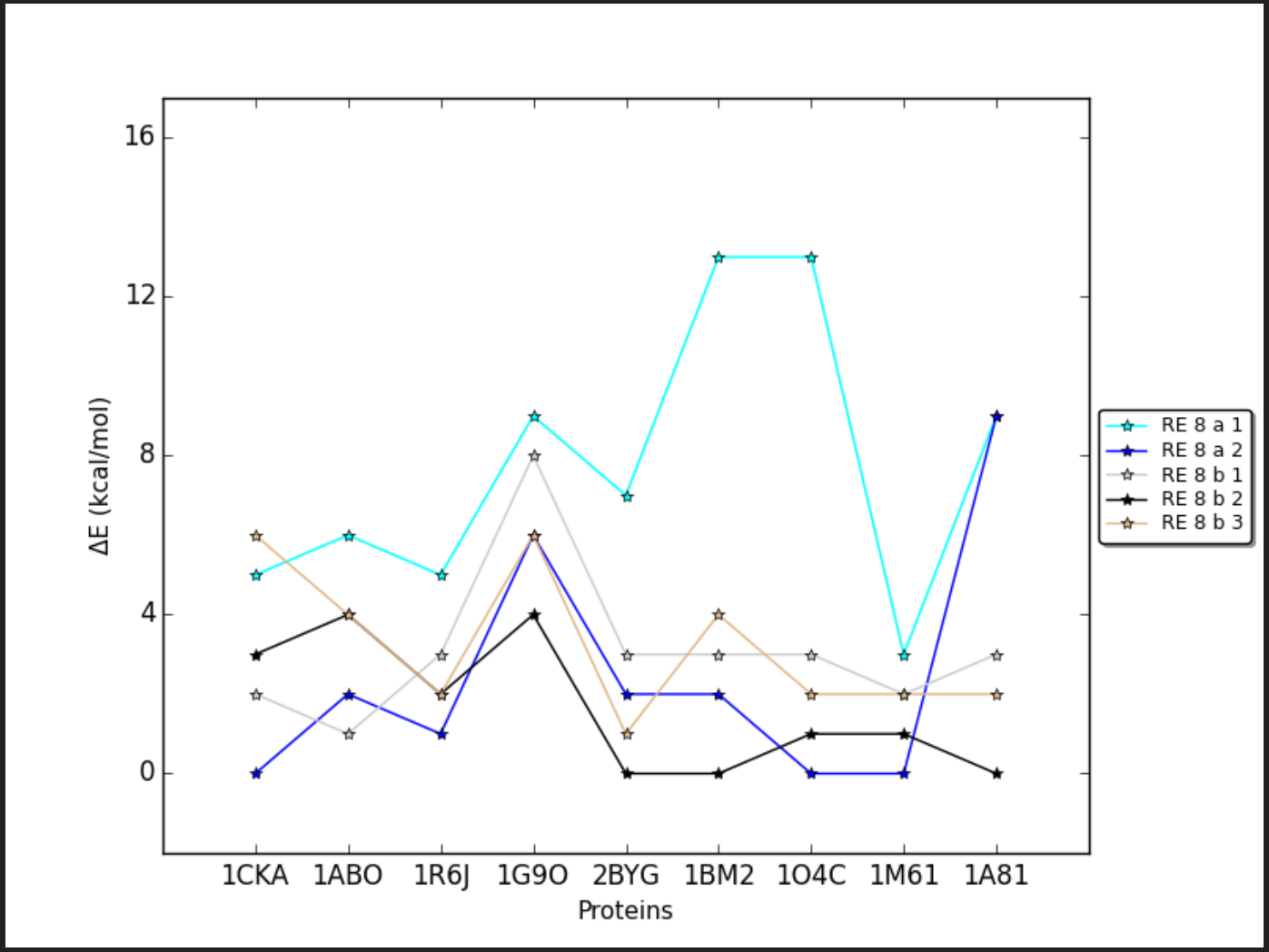
0,1,5,10,20 et 30 positions actives



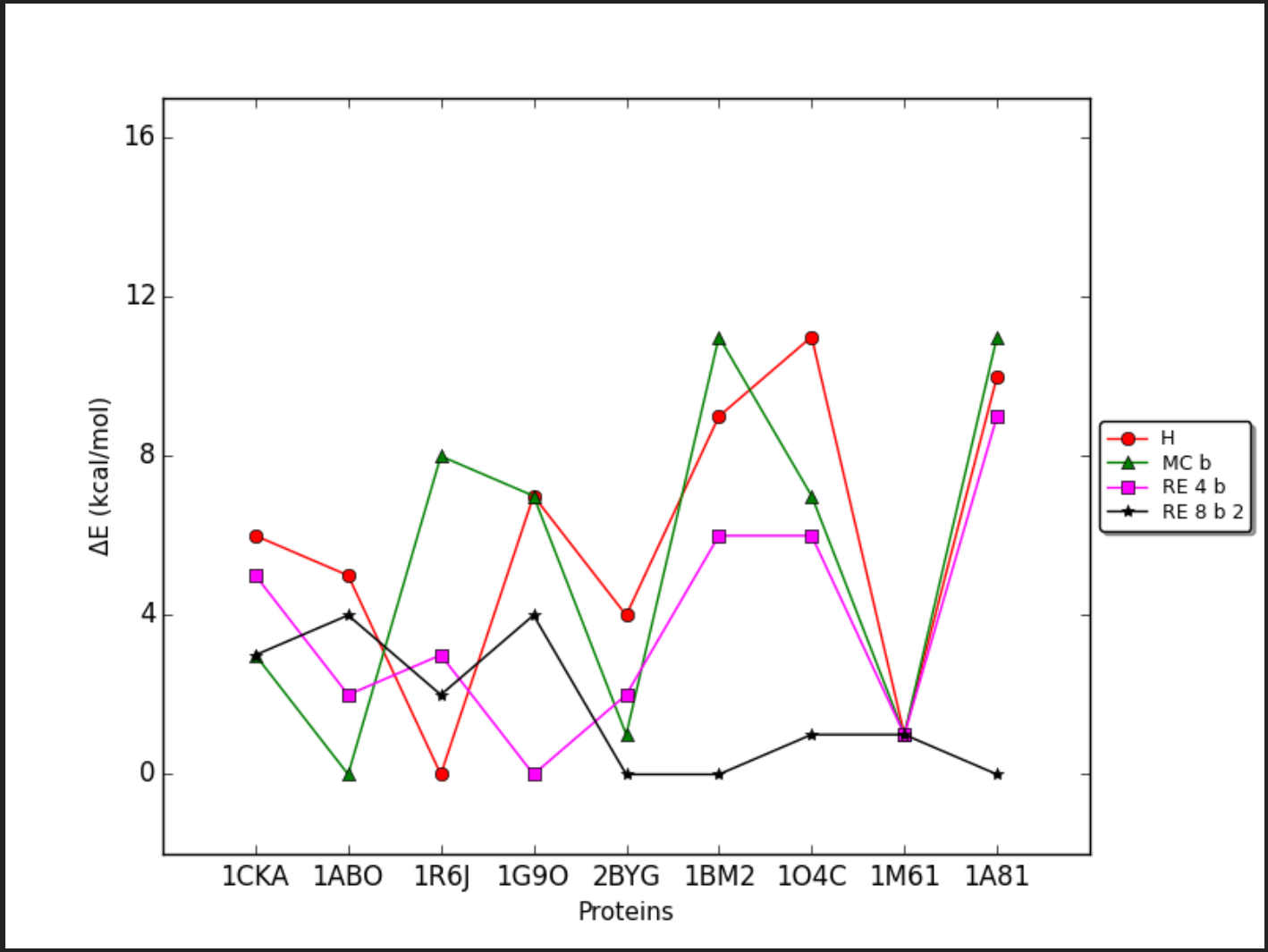
Toutes les positions actives



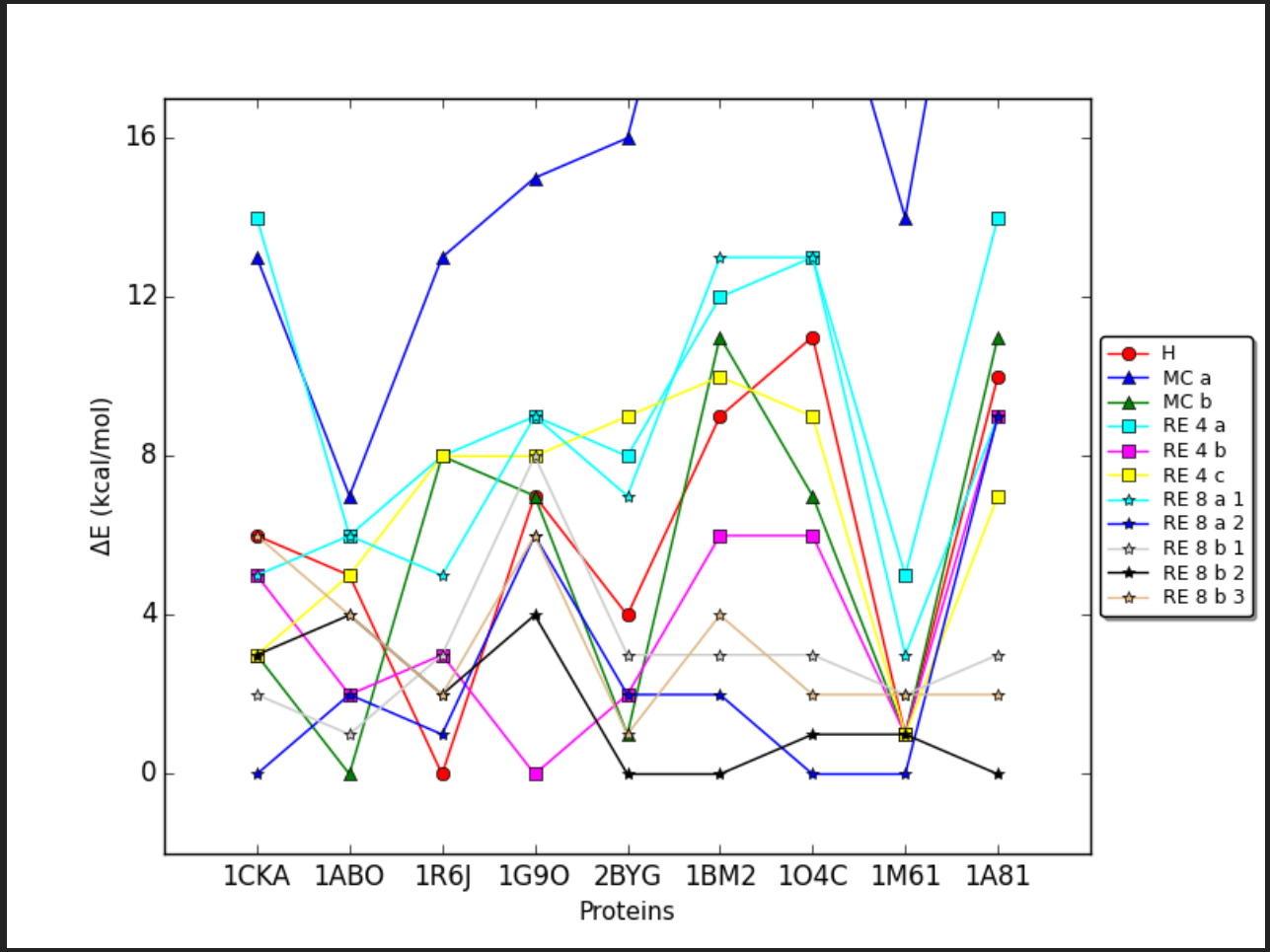
Toutes les positions actives



Toutes les positions actives



Toutes les positions actives



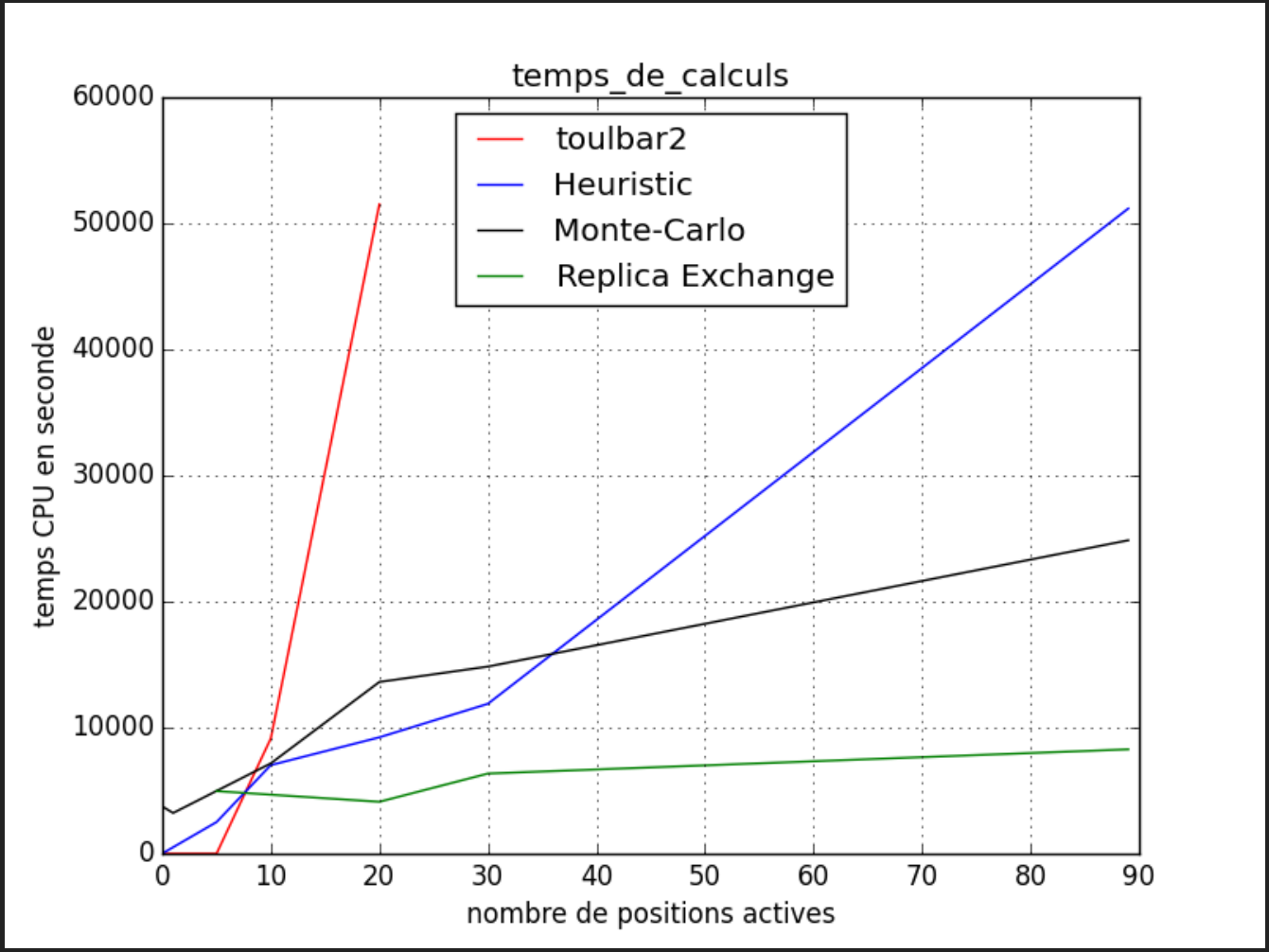
Résultats Superfamily

Pour protocole le RE 8 b 2

Protein	Match/seq size	Superfamily Evalue	superfamily success	Family Evalue	family success
1A81	no				
1ABO	51/58	4.4e-4	100%	2.8e-3	100%
1BM2	78/98	4.2e-5	100%	2.6e-3	100%
1CKA	40/57	1.1e-5	100%	3.4e-3.	100%
1G9O	79/91	7.0e-7	100%	2.5e-3	100%
1M61	97/109	7.2e-7	100%	2.6e-4	100%
1O4C	95/104	2.1e-4	100%	4.5e-3	100%
1R6J	74/82	9.8e-6	100%	4.6e-3	100%
2BYG	59/97	1.4e-5	100%	7.1e-3	100%

Moyenne sur les 10000 séquences-rotamères de meilleurs énergies

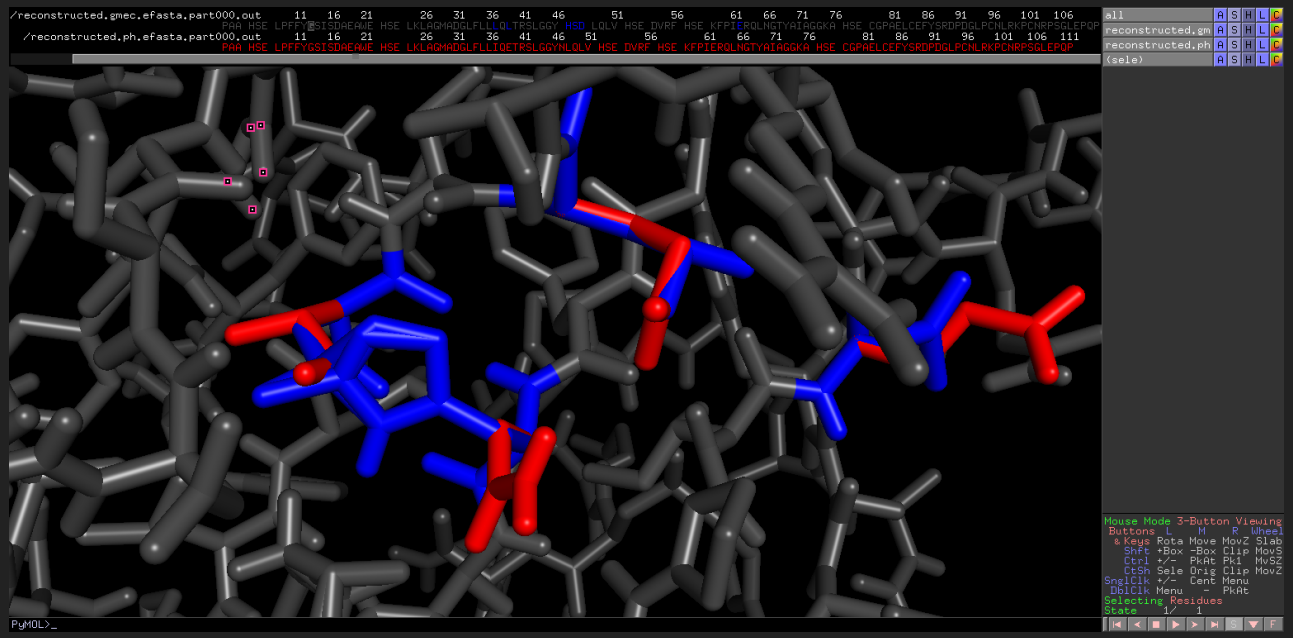
Comparaison des temps de calculs



Retour sur quelques tests

Protein	GMEC	H	MC	RE
1CKA 3	-304.6618	0	0	
1CKA 4	-302.4894	0	0	
1CKA 5	-299.2329	-0.2859	-3.2525	0
1G9O 3	-477.2503	-0.1366	0	
1G9O 4	-470.6458	0	0	
1G9O 5	-464.8659	0	-3.9599	0
1M61 1	-550.0699	0	-0.0776	
1M61 2	-538.6026	-3.5105	-4.5062	0.3215
1M61 5	-553.6559	0	-0.0432	

Protein	seq-rot nb gmec+1	H rank	MC rank	seq nb gmec+1	H mut nb	MC mut nb
1CKA 3	67669	1	1	227	0	0
1CKA 4	4649	1	1	498	0	0
1CKA 5	1388	78	?	77	0	2
1G9O 3	354559	23	1	63	1	0
1G9O 4	22639	1	1	381	0	0
1G9O 5	8658395	1	?	11	0	0
1M61 1	11199153	?	?	21	3	7
1M61 2	11199153	1	1	88	0	0
1M61 5	16417604	1	1	83	0	0



Retour sur quelques tests

Séquences au voisinage du minimum global