Theor Chem Acc (2001) 106:152–157
DOI 10.1007/s002140000243

Theoretical
Chemistry Accounts

Regular article

Exploring a hybrid space minimization procedure

P. Tufféry

INSERM U436, Université Paris 7, case 7113, 2 place Jussieu, 75251 Paris, France
e-mail: tuffery@urbb.jussieu.fr

Received: 13 June 2000 / Accepted: 21 September 2000 / Published online: 21 March 2001

Abstract. Exploring the hypersurface of the energy landscape of proteins remains largely limited owing to the local minimum problem. We present a new hybrid space minimization procedure (HSMP) that couples a side-chain combinatorial search in the rotamer space with a classic minimization procedure in the full dihedral space including backbone variables. The aim of this approach is to enhance the robustness of the overall minimization process by avoiding some of the local minima conditioned by the molecular gear formed by the side chains. The results show, for series of test cases, that lower energies are obtained using HSMP compared to a simple minimization. Perspectives for using such an approach are presented.

Key words: Energy minimization – Multiple minima problem – Proteins

1 Introduction

Statistical mechanics and computer simulation are being used to gain an understanding of features of protein folding or physicochemical properties. A major obstacle in the computation of protein structures is the multiple-minima problem that arises from the existence of many local minima in the multidimensional energy landscape of the protein. Several attempts have been made to overcome this limitation, ranging from limiting the number of variables describing the system (this can be achieved by switching from Cartesian coordinates to internal coordinates or by using simplified models to describe a polypeptidic chain [1]) to focusing on improving the search efficiency of the algorithms employed. The latter paradigm has led to methods such as numerous Monte Carlo/simulated annealing variants [2–6] or genetic algorithms [7, 8] . Attempts to modify the energy functions that simplify the energy landscape of the protein or better fit the native structure of the proteins have also been reported [9–12]. Here, we explore a simple protocol based on the idea that most of the complexity of the energy landscape of proteins is associated with the tight packing of the side chains. Thus, in the case of energy minimization, it is mostly side-chain conformations that will be optimized, while the backbone will mostly undergo the consequences of the modification of the side-chain conformations. To overcome part of this unbalanced process, a classical minimization procedure, performed here in the dihedral space of the protein, is coupled with conformational sampling of the side chains performed in the rotamer space. The algorithm thus evolves in a hybrid space. The sampling of the side chains is expected to allow the minimization to overcome some barriers that in turn will result in a better balance for the backbone optimization.

2 Materials and methods

2.1 Energy calculation

The calculation of the energies of the proteins was achieved using the Flex force field [13, 14]. This force field uses the dihedral space representation and combines classical energy components:

$$E = E_{\text{tor}} + E_{\text{elec}} + E_{\text{vdW}} + E_{\text{Hbonds}} \ ,$$

where E_{tor} is the energy of torsion associated with dihedral angles, E_{elec} the energy associated with the electrostatic interactions, E_{vdW} the nonbonded van der Waals interactions energy, and E_{Hbonds} the energy associated with hydrogen bonds.

The calculations were performed in vacuo, and we used a distance-dependent representation for the dielectric constant of the electrostatics contribution:

$$E^{i,j}_{\text{elec}} = Q_i Q_j / \epsilon_R R_{i,j} \quad \text{where}$$

$$\epsilon_R = D - 0.5(D - D_0)(RS^2 + 2RS + 2)\exp(-RS) \ .$$

The values used here are $D = 78$, $D_0 = 4$, and $S = 0.356$.

2.2 Energy minimization

Energy minimization was performed using the N1QN3 minimizer [15]. It belongs to the variable storage quasi-Newton class of mini-

Contribution to the Symposium Proceedings of Computational Biophysics 2000

mizers and uses the BFGS formula. In such a minimization, the minimizer iteratively estimates the inverse of the Hessian (H^{-1}) to guide the search. However, the estimation of H^{-1} is not accurate until a number of iterations have been performed. Until then, the minimization is closer to a steepest descent. It is important to note is that N1QN3 allows a warm restart (i.e. restarting after interruption).

2.3 Side-chain cluster conformational sampling

The sampling of the side chains was performed using the rotamer approximation to limit the size of the search. Here, the library size was of 602 rotamers to describe the 20 amino acids except glycine and alanine, for which no dihedrals involving heavy atoms exist, and proline, for which the conformation is coupled to that of the backbone. This catalogue corresponds to a 214-rotamer library [16] supplemented by subconformations generated according to the variance associated with each rotamer. Given one side chain, a cluster was defined as the side chains that surround it, using a criterion of distance between the geometric center of a sphere including all possible rotamers for a side chain. Usually, the size of the clusters is less than 6, which allows exhaustive sampling amongst all the possible combinations of rotamers describing the conformations of the side chains of the cluster and the evaluation of the energy associated with each. Here, we are only looking for the combination of rotamers associated with the lowest energy.

2.4 Hybrid space minimization procedure

The hybrid space minimization procedure (HSMP) is as reported in Fig. 1.

At given steps, the minimization is stopped and a side-chain conformational sampling cycle is performed. For amino acids satisfying some triggering condition, a cluster of side chains is defined and a combinatorial search in the rotamer space is performed. On exiting the sampling, the new conformation is accepted if its energy is lower than that of the starting conformation. Once all the clusters are sampled, the gradients are recomputed and the minimization restarted.

Several critical parameters are

- The iteration at which side-chain sampling starts.
- The step of the sampling process, i.e. the number of iterations between two samplings.
- The stopping criterion for the sampling process. Here, we use a self-learning criterion to stop side-chain sampling. It depends on a count of the number of times no gain was obtained for the last samplings.

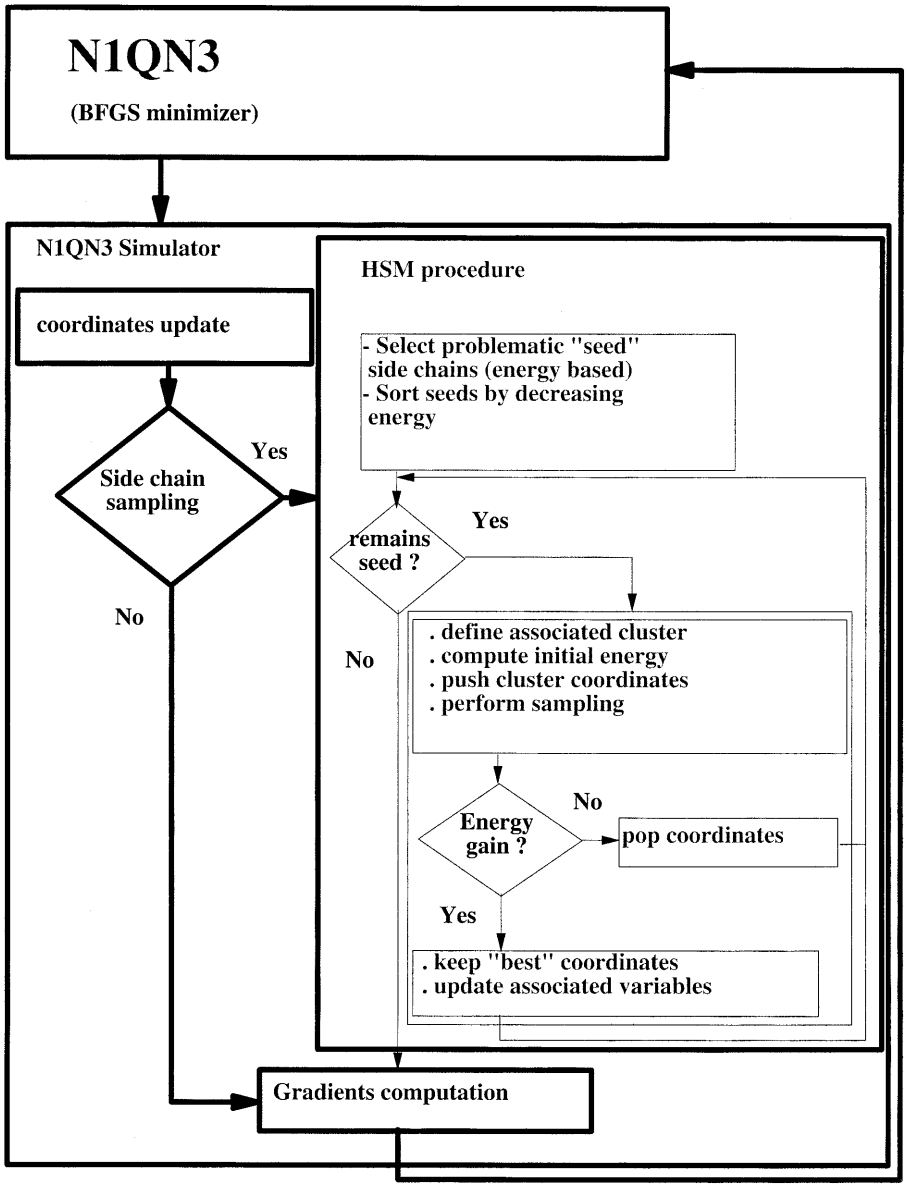

Fig. 1. Flowchart of the hybrid space minimization (HSM) procedure

– The trigger of the sampling centered on a given side chain. In the present study, we use an energy threshold. For one side chain, the existence of energy contacts larger than the threshold will trigger the sampling.

2.5 Selection of proteins and generation of neighbor conformations

Several proteins were chosen from the Protein Data Bank. For each protein, several different conformations were generated. For monomers, they were generated using a random perturbation of the backbone (only ϕ and ψ angles were affected). The perturbation was obtained as the results of numerous (1000) small deviations (usually close to 1°) applied to randomly selected backbone angles. Some control was performed on the root mean square (rms) deviation of the backbone, as well as on the energy deviation relative to the crystallographic conformation. This was done in order to avoid too unrealistic conformations (i.e. no conformations exhibiting crossing backbones or backbone–backbone steric conflicts can be selected). For dimers, the relative orientation of the monomers was modified graphically, using XmMol [17]. For each backbone conformation, the side chains were repositioned, as in a modeling case. In the present work, we used the SMD procedure [18]. For dimers, the backbone was kept rigid during the minimization.

Table 1. Proteins used for the analysis. Entry code: the Protein Data Bank code of the crystallographic structure of the protein. Size: size in amino acids

Protein (abbreviation)	Entry code	Size
Ubiquitin-conjugating enzyme	1aak	150
Guanylate kinase	1gky	188
Lambda repressor	1lmb	179
Elafin/porcine pancreatic elastase complex	1fle	276
Idiotope–anti-idiotope fab–fab complex	1cic	223

3 Assessment of the efficiency of the HSMP

3.1 Procedure calibration

First, one notes that, despite modifying the values of the variables externally to the minimizer, the HSMP does not affect the ability of N1QN3 to converge. For this, the warm restart facility of N1QN3 has to be used after side-chain sampling. Not doing so leads in some cases to premature termination of the minimization, which is not surprising since affecting the variable values also disrupts the search for the descent direction. Using a warm restart, we force the first descent direction to be $-H_{\text{step}}^{-1} g_{\text{step}}$, avoiding such a trap. Reaching convergence using the HSMP suggests that it does not lead to any dramatic inconsistency with H^{-1}.

The sensitivity of the HSMP to some parameters is reported in Table 2 for three datasets. The best results were obtained when starting the HSMP from iteration 10, each five iterations, and with an energy threshold to trigger conformational sampling of 1 kcal/mol (column 4). We will discuss other results taking those as a reference.

Table 2. The sensitivity of the hybrid spare minimization procedure (*HSMP*)

BB_{conf}	ΔE_5^5 1	ΔE_1^1 2	ΔE_1^{5*} 3	ΔE_1^5 4	ΔE_{BB} 5	ΔE_{SB} 6	ΔE_{SS} 7	Δ_{Stp} 8	Δ_{CPU} 9
1aak-tf	−45.85	−24.6	−88.5	−108.1	−50.3	−56.4	1.4	1031	2552
1aak-05	−29.31	41.1	−81.4	−96.5	−8.2	−51.2	−37.1	−402	−282
1aak-10	32.84	78.9	−11.4	−53.3	−44.9	22.5	−30.9	−1262	−1671
1aak-15	38.63	37.3	−47.8	−59.7	3.9	−36.6	−27.0	1239	3173
1aak-20	−48.99	−5.4	−153.3	−151.6	−22.4	−91.8	−37.4	−322	−188
1gky-tf	204.17	67.8	13.5	−52.3	−24.5	−16.5	−11.3	−1708	−4429
1gky-05	102.35	364.2	−12.6	−44.1	5.3	−55.5	6.1	886	3197
1gky-10	101.69	145.8	10.3	−34.4	−5.6	−31.0	2.2	662	2268
1gky-15	216.43	397.2	1.37	−49.1	−1.6	−49.8	2.3	1071	3293
1gky-20	161.61	307.9	−39.0	−46.0	−2.8	−30.6	−12.6	−2154	−4064
1lmb-tf	−53.08	−19.9	−68.0	−72.7	−17.3	−76.8	21.4	−1246	−5839
1lmb-05	−30.44	−3.6	−30.6	−22.8	10.0	−37.6	4.8	−78	186
1lmb-10	−27.53	−0.1	−43.7	−7.8	−3.9	−28.2	24.3	−304	−1067
1lmb-15	−74.91	−55.3	−106.5	−122.0	−23.4	−93.1	−5.5	12	2006
1lmb-20	−12.97	29.4	31.1	−42.4	9.63	−85.8	33.97	−1011	−4248
1fle-1	–	–	–	−129.8	3.6	−65.8	−67.6	66	1473
1fle-2	–	–	–	−113.5	−0.8	−58.3	−54.4	34	522
1fle-3	–	–	–	−195.8	2.7	−109.3	−89.2	12	1869
1cic-1	–	–	–	−178.8	−0.6	−77.8	−100.4	20	2549
1cic-2	–	–	–	−131.1	−0.3	−82.1	−48.7	−54	2751
1cic-3	–	–	–	−141.2	−0.6	−62.7	−77.9	109	2665

BB_{conf}: starting backbone conformation (see Sect. 2)
ΔE_5^5: energy difference obtained with the HSMP starting iteration 10, unlimited, with a step of five iterations, using a threshold of 5 kcal/mol
ΔE_1^1: energy difference obtained with the HSMP starting iteration 10, unlimited, with a step of five iterations, using a threshold of 1 kcal/mol
ΔE_1^5: energy difference obtained with the HSMP starting iteration 10, unlimited, with a step of five iterations, using a threshold of 1 kcal/mol
ΔE_1^{5*}: identical to ΔE_1^5, but with the HSMP limited to 50 first iterations
ΔE_{BB}, ΔE_{SB}, ΔE_{SS}: energy decomposition of ΔE_1, as backbone–backbone (BB), backbone–side chain (SB) and side chain–side chain (SS)
Δ_{Stp}: HSMP number of iterations at convergence, compared to standard minimization (for ΔE_1). Reference is standard minimization
Δ_{CPU}: HSMP computing time (seconds) at convergence, compared to standard minimization (for ΔE_1)

3.1.1 Efficiency of consecutive HSMP cycles

As shown in Fig. 2, most of the efficiency of the HSMP is reached in the early steps of the minimization. Since we are coupling side-chain conformational sampling and minimization, this can be interpreted as a combination of two effects. First, the conformational sampling is per se less and less efficient since it selects side-chain conformations of better and better energy. Second, the energy reference is lower and lower as the minimization process goes along.

Could the late conformational samplings be skipped ? The results obtained when forcing the HSMP to stop at iteration 50 are reported in column 3. Except for some cases where energies can be considered as identical (1aak-20, 1gky-15), the energy values obtained can be largely affected by stopping the HSMP too early. For example, the energy search differs by as much as 40 kcal/mol for 1aak-10, while the sum of the gains of the cycles performed after iteration 50 (column 4) is 20 kcal/mol. Thus, late samplings still seem to allow some better scrutiny of the energy landscape.

3.1.2 Influence of the HSMP step

How does the HSMP perform if sampling is performed at each step of the minimization? As shown in column 2, this leads to somewhat catastrophic results, the energies obtained being in many cases much worse than those obtained by standard minimization (positive values). A possible interpretation is that in such a case the consecutive warm restart cannot allow a correct estimation of H^{-1}. Several tests (not reported) seem to indicate values of the step of the order of five iterations as good values to ensure both efficient side-chain sampling and the preservation of a correct minimization.

3.1.3 Influence of the energy threshold

The results obtained using a threshold of 5 kcal/mol instead of 1 kcal/mol are reported in column 1. Compared to the results obtained for a threshold of 1 kcal/mol, worse results are obtained in all cases, except for 1lmb-05 and 1lmb-10. For many cases, the HSMP also leads to final energy values higher than those obtained by standard minimization (positive values). Possible causes could be that too few cycles of conformational sampling are performed owing to no side-chain selection for late cycles or that too few side chains are selected to undergo the conformational sampling, leaving clusters trapped into local minima in the early iterations. In fact, the overall number of cycles of conformational sampling is not systematically lower than for the 1 kcal/mol threshold. For example, the HSMP is stopped at iteration 139 (or 94) compared to 149 (or 114) for 1aak-10 (or 1aak-05). However, the gain reached by side-chain sampling after iteration 25 is only 17 kcal/mol, compared with 42 kcal/mol obtained with a threshold of 1 kcal/mol. Also, the number of clusters of side chains examined is much lower: 978 versus 2329 for 1aak-05, 1156 versus 2934 for 1aak-10, or 1147 versus 2613 for 1gky-15. Thus, it seems important that the conformational sampling includes a large fraction of side chains.

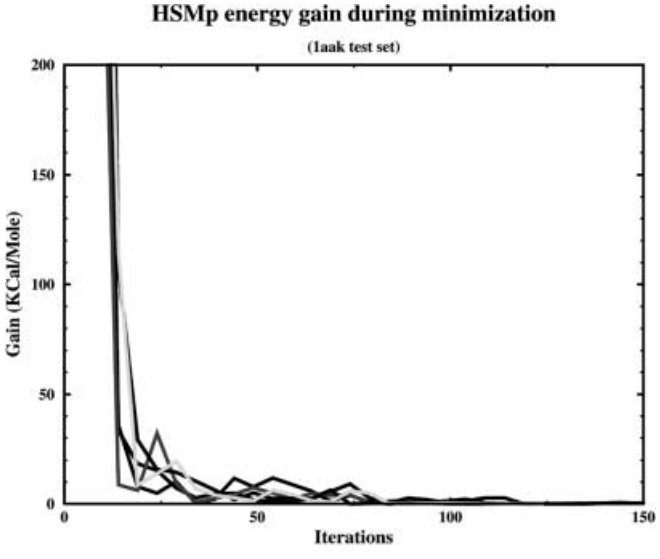

Fig. 2. Gain resulting from side-chain conformational search as a function of the iteration number

3.2 Procedure cost

Since the procedure affects the conditioning of the minimizer through its evaluation of H^{-1}, as well as the search for the descent direction, one can wonder whether the HSMP will not result in much longer minimizations. As shown in Table 2 (column 8), the use of the HSMP does not modify dramatically the number of steps of the minimization. In most cases, it is rather close to the number of steps used for the standard minimization; however, it can be much longer (1aak-15 4000 steps versus 2761, 1cic-3 194 steps versus 85). In some other cases, it can result in much shorter minimizations (1gky-20 3682 steps versus 5836, 1aak-10 2372 steps versus 3634).

Since the procedure requires additional energy computation during side chain sampling, we also report in column 9 the differences in the computing time induced by the HSMP. Again, no obvious tendency can be derived for the monomeric proteins. The difference can be large (close to 1 h), but in some cases, the use of the HSMP leads to important gain as for 1lmb-tf. Different results are obtained for the dimers, for which the use of the HSMP can result in doubling the computing cost. For example, for 1cic-1, we have only a difference of 20 steps of minimization, but more than double the computing cost (2828 s versus 1354 s), for 1671 clusters sampled.

3.3 Improvement in the energy search

Considering our best results (column 4), it is notable that, using such parameterization, the use of the HSMP leads in all cases to conformations of lower energy than those obtained with a classical minimization procedure

(without the HSMP). Also, the order of magnitude of the gain is far from negligible: the smallest gain, obtained for 1gky-10, is still 34.4 kcal/mol.

On analyzing the different components of the energy (backbone–backbone, E_{BB}, backbone–side chains, E_{SB}, side chains–side chains, E_{SS}), one observes for monomeric chains that, in most cases, the best gain is for the backbone–side chain component, while the side chain–side chain or backbone–backbone energy differences can be unfavorable in some cases. Hence, a major effect of the procedure seems to be its ability to solve conformational maladjustments between backbone and side chains. This makes sense since when positioning side chains prior to the minimization, no backbone conformational flexibility was considered at this stage. Thus, a possible interpretation is that, in the rotamer space and for a fixed backbone conformation, while side chain–side chain conflicts can be reasonably solved, side chain–backbone conflicts are much worse solved, since only some conformational flexibility of the side chain conformation and none of the backbone can be used.

For the dimers, the goal of the present tests was to assess whether the HSMP could be efficient for semi flexible docking, and the backbone of the monomers was locked. Hence E_{BB} represents only the difference in energy due to the relative positions of the individual monomers. The balance between E_{SB} and E_{SS} appears more equilibrated.

3.4 Induced modification of the protein conformation

The C_α rms deviations induced by the different minimization procedures (columns 1 and 2) and the C_α rms deviations between the final conformations are reported in Table 3. rms deviations induced by the standard and HSMP minimizations are close for most cases. For 1aak-tf, 1lmb-tf, 1lmb-15, and 1lmb-20, the HSMP leads to smaller deviations, while for 1aak-15, 1lmb-10, and 1cic-3 the HSMP leads to larger deviations. However, comparing the final conformations shows that the minimization procedures lead to conformations that can be rather different, with deviations as large as 4.21 Å for 1aak-15. In this case, however, most of the deviation comes from the N-terminal extremity. By fitting the structures removing this loop, the rms deviations decrease to 1.84 Å for the standard minimization and 1.47 Å for the HSMP. As shown in Fig. 3, the resulting conformation of the HSMP is much closer to the starting one, the major differences being located in the secondary structure junctions, while the deviation is more widely spread for the conformation resulting from the standard minimization. Backbone–backbone energies are similar (Table 2, column 5). Hence, not using the HSMP, the backbone of the protein seems to undergo larger modifications as a means to solve the energy descent, while increased side-chain flexibility resulting from the HSMP does not require it.

Table 3. Root-mean square deviations (rmsd)

BB_{conf}	$rmsd^1$ 1	$rmsd^2$ 2	$rmsd_1^2$ 3	$\%_{\chi 1-2}$ 4	$\%_{\chi 1-2}^{HSMP}$ 5
1aak-tf	3.62	1.75	3.57	90	57
1aak-05	1.80	1.86	2.38	85	55
1aak-10	2.42	2.23	2.82	82	55
1aak-15	2.29	3.71	4.21	85	49
1aak-20	3.45	3.43	3.00	85	47
1gky-tf	1.93	2.54	1.77	96	76
1gky-05	2.28	2.25	1.22	95	66
1gky-10	2.93	2.76	0.97	93	76
1gky-15	2.05	2.06	1.45	94	75
1gky-20	2.79	2.97	1.60	90	69
1lmb-tf	3.96	2.39	2.26	96	62
1lmb-05	2.09	2.87	1.73	97	72
1lmb-10	2.64	2.97	1.54	96	71
1lmb-15	3.10	2.25	2.18	92	63
1lmb-20	4.06	2.51	2.68	96	65
1fle-1	0.29	0.42	0.32	95	79
1fle-2	0.41	0.55	0.46	97	81
1fle-3	0.78	0.68	0.89	96	87
1cic-1	0.84	1.07	0.40	97	80
1cic-2	0.78	0.98	0.55	96	85
1cic-3	1.27	1.65	1.34	95	74

BB_{conf}: starting backbone conformation (see Sect. 2)
$rmsd^1$: C_α rmsd (Å) between starting and final conformations, using standard minimization
$rmsd^2$: C_α rmsd (Å) between starting and final conformations, using the HSMP minimization (for ΔE_1)
$rmsd_1^2$: C_α rmsd (Å) between standard and the HSMP minimized protein conformations
$\%_{\chi 1-2}$ (or $\%_{\chi 1-2}^{HSMP}$): fraction of side chains for which final χ_1 and χ_2 are not different by more than 40° from initial ones (standard minimization, or the HSMP)

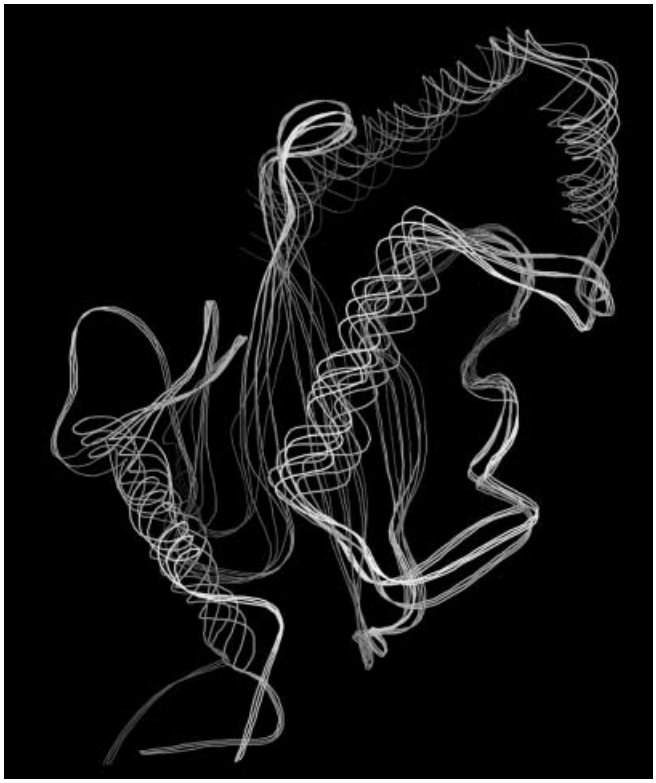

Fig. 3. Superimposed structures of 1aak-15 onto the crystallographic conformation (1aak). *Pink*: conformation obtained by standard minimization. *Yellow*: conformation obtained using the HSM procedure

In terms of side-chain conformations, columns 4 and 5 show that using the HSMP results in much more conformational change than the standard minimization, which was expected. While the latter preserves more than 90% of χ_1 and χ_2, the former modifies as much as 30% of them.

4 Conclusions

The HSMP procedure was designed with the aim of allowing a better energy search and a better balance between the optimization of the side-chain and main-chain conformations. Our results show that the correct parameterization of such a procedure can indeed lead to the quasi-systematic energy gain compared to the standard minimization, with nonsignificant cost in terms of the number of iterations or computing cost for monomers. In terms of conformations, the use of the HSMP leads to rather different backbone conformations and allows more side-chain conformational changes.

Such a procedure, tested here in the case of a simple minimizer could be embedded in more sophisticated search procedures, such as genetic algorithms or Monte Carlo techniques. However, reaching the experimental conformation also poses the problem of the accuracy of the force field employed.

References

1. Lee J, Liwo A, Ripoll DR, Pillardy J, Scheraga HA (1999) Proteins Suppl 3: 204–208
2. Li Z, Scheraga HA (1987) Proc Natl Acad Sci USA 84: 6611–6615
3. Shin JK, Jhon MS (1991) Biopolymers 31: 177–185
4. Caflisch A, Neiderer P, Anliker M (1992) Proteins 14: 102–109
5. Scheraga HA (1996) Biophys Chem 59: 329–339
6. Trosset JY, Scheraga HA (1998) Proc Natl Acad Sci USA 95: 8011–8015
7. Rabow AA, Scheraga HA (1996) Protein Sci 5: 1800–1815
8. Pedersen JT, Moult J (1997) J Mol Biol 269: 240–259
9. Purisima EO, Scheraga HA (1987) J Mol Biol 196: 697–709
10. Hao MH, Scheraga HA (1996) Proc Natl Acad Sci USA 93: 4984–4989
11. Hao MH, Scheraga HA (1999) Curr Opin Struct Biol 9: 184–188
12. Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA (1999) Proc Natl Acad Sci USA 96: 5482–5485
13. Lavery R, Sklenar H, Zakrzewska K, Pullman B (1986) J Biomol Struct Dyn 3: 989–1014
14. Lavery R, Parker I, Kendrick J (1986) J Biomol Struct Dyn 4: 443–462
15. Gilbert JC, Lemaréchal C (1989) Math Program 45: 407–435
16. Tuffery P, Etchebest C, Hazout S (1997) Protein Eng 10: 361–372
17. Tuffery P (1995) J Mol Graph 13: 67–72
18. Tuffery P, Etchebest C, Hazout S, Lavery R (1991) J Biomol Struct Dyn 8: 1267–1289