Eur Biophys J (2007) 37:95–104
DOI 10.1007/s00249-007-0170-2

ORIGINAL PAPER

A simple genetic algorithm for the optimization of multidomain protein homology models driven by NMR residual dipolar coupling and small angle *X*-ray scattering data

Fabien Mareuil · Christina Sizun · Javier Perez · Marc Schoenauer · Jean-Yves Lallemand · François Bontems

Received: 12 February 2007 / Revised: 10 April 2007 / Accepted: 21 April 2007 / Published online: 24 May 2007

Abstract Most proteins comprise several domains and/or participate in functional complexes. Owing to ongoing structural genomic projects, it is likely that it will soon be possible to predict, with reasonable accuracy, the conserved regions of most structural domains. Under these circumstances, it will be important to have methods, based on simple-to-acquire experimental data, that allow to build and refine structures of multi-domain proteins or of protein complexes from homology models of the individual domains/proteins. It has been recently shown that small angle X-ray scattering (SAXS) and NMR residual dipolar coupling (RDC) data can be combined to determine the architecture of such objects when the X-ray structures of the domains are known and can be considered as rigid objects. We developed a simple genetic algorithm to achieve the same goal, but by using homology models of the domains considered as deformable objects. We applied it to two model systems, an S1KH bi-domain of the NusA protein and the γS-crystallin protein. Despite its simplicity our algorithm is able to generate good solutions when driven by SAXS and RDC data.

F. Mareuil · C. Sizun · J.-Y. Lallemand · F. Bontems (✉)
ICSN-RMN, Institut de Chimie des Substances Naturelles 91190
Gif-sur-Yvette and Ecole Polytechnique,
91128 Palaiseau, France
e-mail: francois.bontems@polytechnique.fr

J. Perez
SWING, Synchrotron SOLEIL, L'Orme des Merisiers,
Saint-Aubain, BP 48, 91192 Gif-sur-Yvette Cedex, France

M. Schoenauer
LRI, CNRS-Université Paris Sud, 91400 Orsay, France

Keywords Genetic algorithm · Residual dipolar coupling · SAXS · Multi-domain protein · Homology models

Introduction

Built on the success of genome sequencing projects, structural genomic projects aim at collecting extensive data about the relationships between protein sequences, structures and functions. Rapid large-scale protein structure determination has benefited from fast data acquisition and analysis as well as high-throughput screening for protein expression and purification. But structure determination of all gene products is out of reach. In fact, structural genomics rely on the paradigm that the number of protein folds is finite. Along with the principle objectives of the projects, one common objective is to produce representative structures of all sequence families (Todd et al. 2005) with the hope that the satisfaction of this objective will allow the prediction by homology of any protein structure.

This strategy is facing two problems. First, even if a similarity between two sequences as low as 30% is sufficient to ensure that they correspond to similar folds, the quality of the structure prediction falls drastically with the similarity score between the sequences (Ginalski 2006). In addition, modeling of loops remains challenging. Second, so far structural efforts have mainly concentrated on compact globular single domains. But most genes encode for multi-domain proteins. Determination of relative domain arrangement has remained difficult for both X-ray diffraction and nuclear magnetic resonance. On the one hand, the influence of crystal packing forces raises questions about the relevance of quaternary structure determined by X-ray analysis. On the other hand, intermolecular nOe measured between domains are generally scarce and not sufficient to constraint a multi-domain protein structure. Under these circumstances, we believe that it is important to have a disposal of methods that allow the refinement of homology predicted domain structures and the relative positioning of domains in multi-domain proteins from data easily accessible in solutions, in particular from residual dipolar coupling and small angle X-ray scattering data.

Residual dipolar coupling (RDCs) measured in anisotropic solvent (Tjandra and Bax 1997) have prompted a large interest in the NMR community (Bax 2003; Lipsitz and Tjandra 2004; Prestegard et al. 2004). RDCs contain information about the orientations of internuclear vectors with respect to the magnetic field, not accessible by other methods. They can be used to constrain the relative orientations of protein fragments, e.g. peptidic planes, residues, secondary structural elements or domains (in multi-domain proteins). Moreover, they mostly rely on backbone atom resonances, which can be easily assigned even in large perdeuterated proteins, and they require only little measurement time. Accordingly, RDCs have been widely used to refine protein structures in complement of nOe restraints. They have also been shown to be of particular interest for protein/domain docking in conjunction with SAXS data (Mattinen et al. 2002; Grishaev et al. 2005; Gabel et al. 2006). As the SAXS curve is the Fourier transform of the distance distribution function of the complex/protein, it provides translational information about the relative positioning of the proteins/domains that complements the orientational restraints obtained from RDCs. So far, all reported examples have used experimentally determined high-resolution structures of the proteins/domains to be associated. The proteins/domains were considered as rigid and RDC and SAXS information contributed only to positioning. When imperfect homology models are used, the proteins/domains have to be allowed to deform and RDCs and SAXS should also contribute to their refinement. In addition, the interpretation of RDCs in terms of orientation requires the accurate determination of the alignment tensor. Which is not trivial when the starting structures of the proteins/domains are approximate.

In this paper, we present a simple genetic algorithm to calculate bi-domain protein structures, based on homology models of the individual domains and driven by RDC and SAXS data. We tested our procedure on two systems. The first uses only simulated data. The second uses experimental RDCs extracted from the BMRB data bank. We show that when using this tool it is indeed possible to find good models in both cases. We also discuss the limits and the possible improvements of the procedure.

Methods

Organization of the genetic algorithm

We chose a genetic algorithm among various optimization techniques because this inverse method requires only evaluation of an objective function and not of its gradient. In addition, it is easily adaptable and additional information sources can be taken into account by simply modifying the objective function.

Genetic algorithms are stochastic numerical optimization procedure roughly inspired from population genetics principles (Eiben and Smith 2003). A population of candidate solutions (individuals or chromosomes) is submitted to variation and selection cycles, resulting in its enrichment in better and better solutions. The variations result from the application of unary mutation and binary recombination operators. Selection relies on comparative ‘‘fitness’’ (value of the objective function). The detailed scheme used here is presented in Fig. 1. The individuals are encoded as PDB files. The *N* (typically 25) parental structures are first mutated, each leading to *n* (typically five) children. The children are recombined with each another. The fitness of the *n***N* individuals is evaluated and a new set of *N* parents is selected by a tournament method. To be easily modifiable and portable, the algorithm is written as a C-shell that manages variables and files, and launches the modification, evaluation and selection procedures. Mutation and recombination are realized by X-PLOR scripts (Brünger 1992), RDC and SAXS evaluations by a home-written SVD procedure and by the Crysol software (Svergun et al. 1995) respectively, and selection by a python script. All procedures communicate through ASCII files.

Mutation

Inside the mutation procedure, cartesian coordinates are converted to internal coordinates, edited, and converted back to cartesian. After minimization of its side-chain energy, a mutated structure is accepted or rejected according to a predefined energy threshold. The procedure is repeated until the correct number of mutations per child and children per parent is obtained. Two kinds of mutations can be applied. A global mutation consists of a random perturbation of the backbone torsion angles (φ,ϕ) of a residue, introducing a displacement of all following residues. A local mutation induces the perturbation of a small portion of the protein chain without modifying the rest of the structure. Like in a global mutation, the (φ,ϕ) torsion angles of a residue are perturbed, but the resulting structure is recombined with its parent three to five residues downstream and the geometry of the modified segment is

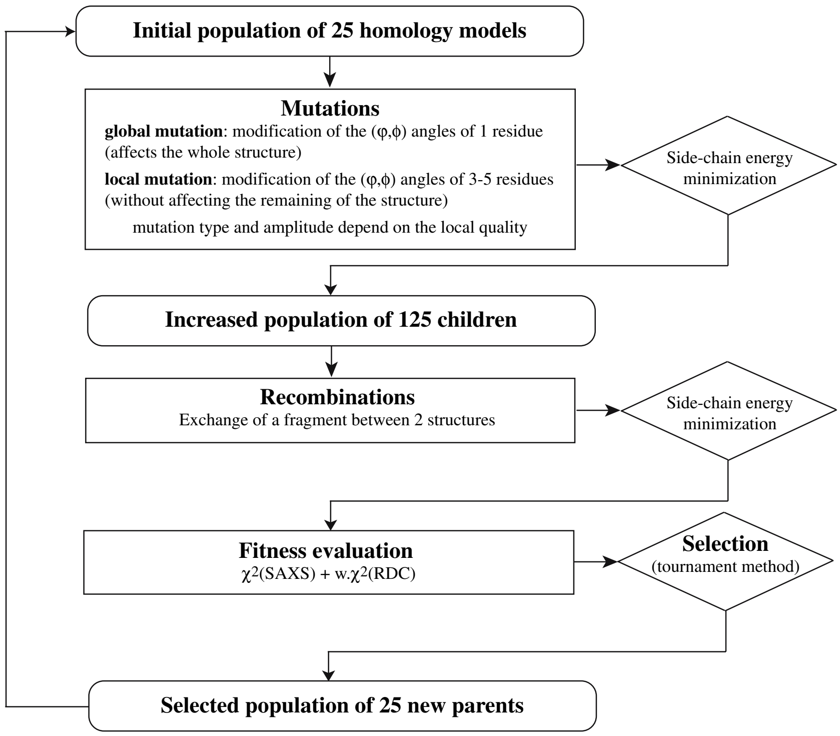

Initial population of 25 homology models
Mutations
global mutation: modification of the (φ,ϕ) angles of 1 residue (affects the whole structure)
local mutation: modification of the (φ,ϕ) angles of 3-5 residues (without affecting the remaining of the structure)
mutation type and amplitude depend on the local quality
Side-chain energy minimization
Increased population of 125 children
Recombinations
Exchange of a fragment between 2 structures
Side-chain energy minimization
Fitness evaluation
χ2(SAXS) + w.χ2(RDC)
Selection
(tournament method)
Selected population of 25 new parents

Fig. 1 Flow chart of the genetic algorithm

repaired by a simple minimization. The mutation kind, position and amplitude are adapted to the structural context and to the local quality of the structure. Three types of regions are defined in the protein and encoded in the B factor. The well predicted regions in the initial homology models, typically secondary structure elements, are encoded by B = 1 (first domain) and B = 2 (second domain). The poorly predicted regions, generally loops, are encoded by B = 0 and the linker by B = –1. Two out of three mutations are introduced in the linker (B = –1); 70% of the remaining in the poorly predicted segments (B = 0) and 30% in the well predicted regions (B = 1 or 2). In the linker, all mutations are global, The local agreement between the structures and the experimental data is encoded in the *Q* factor. The (φ,ϕ) angle perturbations are generated by choosing a Gaussian random number with 0 mean and standard deviation defined by *K**<*Q*>reg/<*Q*>prot where *K* is a small predefined value, <*Q*>reg is the mean *Q* factor averaged over the perturbed region, and <*Q*>prot is the mean *Q* factor averaged over the whole structure.

Recombination

In the recombination scripts, the well-predicted regions of one domain (B = 1 or B = 2) of two parents are superimposed. Two recombination points are chosen randomly in the superimposed domains, the fragments are exchanged and the side-chain energy is minimized. The procedure is repeated until an acceptable structure (residual energy lower than a threshold) is obtained. A residue is eligible as recombination point when it belongs to a three-residue fragment for which the distance between the backbone atoms of the two parental structures is less than 1.0 Å.

RDC and SAXS evaluation

The fitness function is calculated as the weighted sum of the SAXS and RDC evaluation functions. The RDC evaluation function is the root mean-square deviation between experimental values and those calculated from the molecular alignment tensor **A** determined for each structure:

$$\begin{aligned} \frac{D_{ij}}{D_{ij}^{\max}} &= \frac{1}{2}\left[SA_a\left(3\cos^2\theta - 1\right) + \frac{3}{2}SA_r\sin^2\theta\cos 2\phi\right] \\ D_{ij}^{\max} &= -\frac{\mu_0}{4\pi^2}\hbar\frac{\gamma_i\gamma_j}{r_{ij}^3} \end{aligned} \qquad (1)$$

θ and ϕ are the polar coordinates of an internuclear vector *ij* relative to the principal axis system of **A**. The alignment tensor (axial component A_a, rhombicity $R = A_r/A_a$ and the 3 Euler angles giving the orientation of the molecular frame in the principal axis system) is obtained by a singular value decomposition (SVD) algorithm, as first proposed by Prestegard et al. (Losonczi et al. 1999), adapted from the Numerical Recipes subroutines in C (Press et al. 1992) and optimized by a Monte-Carlo procedure. SVD is only performed on the first module for residues with B = 1. The deviation between experimental and calculated RDCs for each residue is stored in the structural Q-factor. For the SAXS curve, the deviation is calculated according to Eq. 2 (Press et al. 1992). The value was normalized, in the Figs. 3–5, by the result obtained with the target structures.

$$\chi^2 = \sum_i \frac{\left(\sqrt{\frac{S}{R}}R_i - \sqrt{\frac{R}{S}}S_i\right)^2}{R_i + S_i} \text{ with}$$
$$R = \sum_i R_i \text{ and } S = \sum_i S_i \quad (2)$$

Selection

In the first 200 calculation cycles, the objective is to enrich the population with good individuals while preserving large variability. To fulfill this, the selection is based on a tournament method: small groups of individuals (two in our case) are randomly uniformly extracted from the mutated-recombined population and the best of each group survives to the next generation. However, with such a low selection pressure, the procedure generally does not reach a minimum. Accordingly, a second stage of 100 calculation cycles is added with a higher selective pressure: the parental population is regenerated by selecting the best individuals in the mutated-recombined population. In addition, to ensure that this procedure always leads to an improvement of the population, the parents are included in the mutated-recombined population.

Initial data generation

RDC and SAXS data

In the case of γS-crystallin, we extracted the H-N, C^α-C' and N-C' couplings of one of the two RDC constraint sets deposited at the BMRB. In the case of the S1KH bi-module of *T. martima* NusA (*Tm*-S1KH), we generated synthetic N-H, C^α-H^α, N-C' and C^α-C' dipolar couplings from the order parameters published for human ubiquitin (Saa = 8.3 10^{-4} and SAr = 1.4 10^{-4}) to which we added a random noise (10% of the maximal value). In both cases, we calculated a SAXS curve with the CRYSOL software (Q between 0.2 and 5 nm^{-1}), to which we added a realistic random noise following a normalized Gaussian distribution. To stick to what experimental measurement would have given, the width of the noise distribution was taken proportional to $\frac{1}{\sqrt{Q}}\sqrt{(I_{\mathrm{sol}}(Q) + I_{\mathrm{Buf}})}$, where I_{Buf} is the constant scattering intensity from the buffer and $I_{\mathrm{Sol}}(Q)$ (=$I_{\mathrm{Cry}}(Q)$ + I_{Buf}) is the scattering intensity from the protein solution. I_{Cry} (Q) is the curve given by CRYSOL. Realistically, I_{Buf} was taken as ten times the value of $I_{\mathrm{Cry}}(Q)$ at Q = 5 nm^{-1}. The $\frac{1}{\sqrt{Q}}$ factor arises from the fact that a 2D detection is considered.

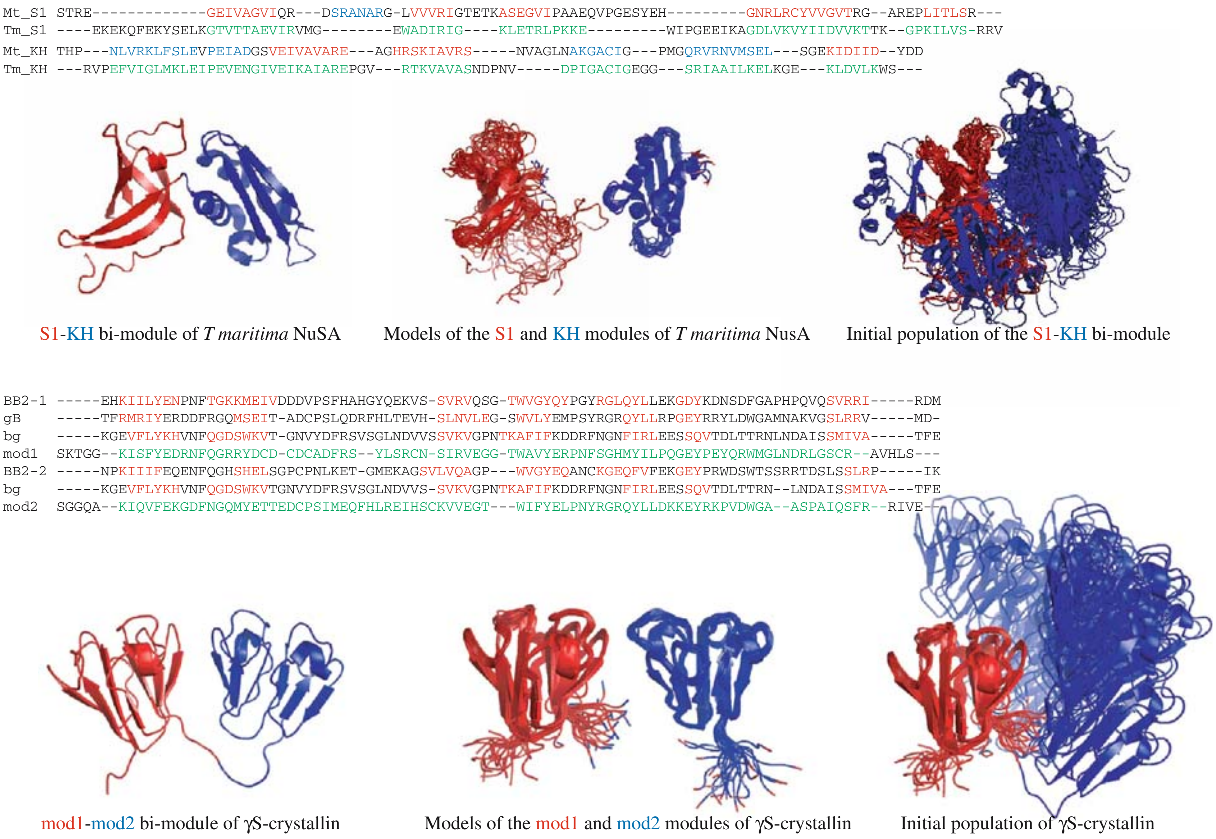
Model building

We needed large families of initial structures as seed for the genetic algorithm. We could have modeled the whole bi-domain, by using the structures of S1KH of *M. tuberculosis* NusA (*Mt*-S1KH) to model *Tm*-S1KH and βB2-crystallin to model γS-crystallin. However, to address a more general situation, we modeled each domain independently. The alignments are presented in Fig. 2. For S1 and KH, we used only one template for each domain (structures of *Mt*-S1 and *Mt*-KH). Accordingly, we only imposed the conservation of the secondary structure elements and left the loops free. In the case of the two γS-crystallin domains, we used two (for domain 2) and three (for domain 1) templates. The template sequences were first aligned on the basis of the superimposition of the template structures. The target sequences were then aligned using the MALIGN routine of the Modeller software (Sali and Blundell 1993). In both cases, the structure of the linker regions was not imposed. The Modeller software was further used to build 1000 models of each domain, which were oriented with respect to the alignment tensor determined from the RDCs of the well-predicted regions (B = 1 or B = 2). A priori, four orientations are possible for each domain, due to the degeneracy of the RDCs. One of these orientations was chosen for S1 and the first γS-crystallin domain, while the four orientations were conserved for KH and the second γS-crystallin domain. Initial bi-domain structures were built by randomly selecting structures for the first and the second domains. The C-terminus of the first domain was brought in the vicinity of the N-terminus of the second by translation and the structure of the linker region was minimized. Structures for which the interaction energy between the two domains was negative or null were conserved. They were then clustered on the basis of the relative orientation of the domains. Only two orientations were possible for S1KH and three for γS-crystallin, the others systematically leading to domain overlap. The final structures (250 for each orientation) were divided into initial populations of 25 individuals that were optimized in parallel.

Results

Starting from the bi-domain structure

We first examined to what extent residual dipolar coupling (RDC) and small angle X-ray scattering (SAXS) data were



```
Mt_S1 STRE-------------GEIVAGVIQR---DSRANARG-LVVVRIGTETKASEGVIPAAEQVPGESYEH---------GNRLRCYVVGVTRG--AREPLITLSR---
Tm_S1 ----EKEKQFEKYSELKGTVTTAEVIRVMG--------EWADIRIG----KLETRLPKKE---------WIPGEEIKAGDLVKVYIIDVVKTTK--GPKILVS-RRV
Mt_KH THP---NLVRKLFSLEVPEIADGSVEIVAVARE---AGHRSKIAVRS-----NVAGLNAKGACIG---PMGQRVRNVMSEL---SGEKIDIID--YDD
Tm_KH ---RVPEFVIGLMKLEIPEVENGIVEIKAIAREPGV---RTKVAVASNDPNV-----DPIGACIGEGG---SRIAAILKELKGE---KLDVLKWS---
```

S1-KH bi-module of *T maritima* NusA   Models of the S1 and KH modules of *T maritima* NusA   Initial population of the S1-KH bi-module

```
BB2-1 -----EHKIILYENPNFTGKKMEIVDDDVPSFHAHGYQEKVS-SVRVQSG-TWVGYQYPGYRGLQYLLEKGDYKDNSDFGAPHPQVQSVRRI-----RDM
gB    -----TFRMRIYERDDFRGQMSEIT-ADCPSLQDRFHLTEVH-SLNVLEG-SWVLYEMPSYRGRQYLLRPGEYRRYLDWGAMNAKVGSLRRV-----MD-
bg    -----KGEVFLYKHVNFQGDSWKVT-GNVYDFRSVSGLNDVVSSVKVGPNTKAFIFKDDRFNGNFIRLEESSQVTDLTTRNLNDAISSMIVA-----TFE
mod1  SKTGG--KISFYEDRNFQGRRYDCD-CDCADFRS--YLSRCN-SIRVEGG-TWAVYERPNFSGHMYILPQGEYPEYQRWMGLNDRLGSCR--AVHLS---
BB2-2 -----NPKIIIFEQENFQGHSHELSGPCPNLKET-GMEKAGSVLVQAGP---WVGYEQANCKGEQFVFEKGEYPRWDSWTSSRRTDSLSSLRP-----IK
bg    -----KGEVFLYKHVNFQGDSWKVTGNVYDFRSVSGLNDVVS-SVKVGPNTKAFIFKDDRFNGNFIRLEESSQVTDLTTRN--LNDAISSMIVA---TFE
mod2  SGGQA--KIQVFEKGDFNGQMYETTEDCPSIMEQFHLREIHSCKVVEGT---WIFYELPNYRGRQYLLDKKEYRKPVDWGA--ASPAIQSFR--RIVE--
```

mod1-mod2 bi-module of γS-crystallin   Models of the mod1 and mod2 modules of γS-crystallin   Initial population of γS-crystallin

Fig. 2 Initial population building. The sequences of the templates (*S1* and *KH* domains of *Mycobacterium tuberculosis*: NusA *Mt*-S1 and *Mt*-KH; first and second domains of βB2-crystallin: *BB2-1*, *BB2-2*; γB-crystallin: gB; $\beta\gamma$-crystallin: bg) are aligned with those of the targets (*S1* and *KH* domains of *Thermotoga maritima*: *Tm-S1*, *Tm-KH*; first and second domain of *γS-crystallin*: *mod1*, *mod2*). The secondary elements of the templates are indicated in *blue* (α-helices) and in *red* (β-strands). The regions constrained in the target modeling are in *green*. The target structures (S1KH bi-module of *T. maritima* and *γS-crystallin*) are displayed on the *left side*. A set of models for the isolated domains is in the middle, and an initial family composed of 25 structures is on the *right side*

able to define the structures to be optimized. To address this question, we first ran a series of calculations starting with the target structures (*Tm*-S1KH and γS-crystallin) and we looked at their deformation after 200 cycles of the genetic algorithm with low selective pressure and after 100 additional cycles with high selective pressure (see ''Methods''). In the case of γS-crystallin, the data set was composed of experimental RDCs and of a simulated SAXS curve with noise In the case of S1KH, we simulated RDC and SAXS data sets with noise.

As illustrated by Fig. 3, we observed a divergence of the structures, reflected by an increase of the backbone rmsd calculated between the ''optimized'' and the target structures. At the end of the first 200 low selective pressure cycles the divergence is significant. The rmsd spread is 1–4 Å for S1KH and 1–6 Å for γS-crystallin. If the results are restricted to the structures with the best fitnesses, the spread is less, but there are still equivalent solutions up to 3 Å. It should be noted that the fitnesses at the end of this stage are always much larger than those determined for the targets (43 for S1KH, 104 for γS-crystallin). The SAXS curve, in particular, is never correctly fitted with the low selection pressure. This is not surprising if we consider that low pressure was deliberately chosen to favor variability among the populations. The situation is improving after additional 100 high selective pressure cycles. The fitnesses remain larger than those of the targets, but much better solutions are retrieved. The rmsd spreads are now 1.5–2.5 Å for S1KH and 1.7–4 Å for γS-crystallin. Moreover, the two lowest fitness S1KH families (fitness = 129 and 132) have rmsds calculated over the family of 1.51 and 1.71 Å. Similarly, the four lowest fitness γS-crystallin families (fitness = 202, 220, 224 and 230) have rmsds of 1.70, 1.79, 1.44 and 1.59 Å respectively.

Starting with perfect domain structures

In a second step, we tested to what extent we were able to determine the structure of a bi-module from the crystallographic structures of the domains. This time, we built

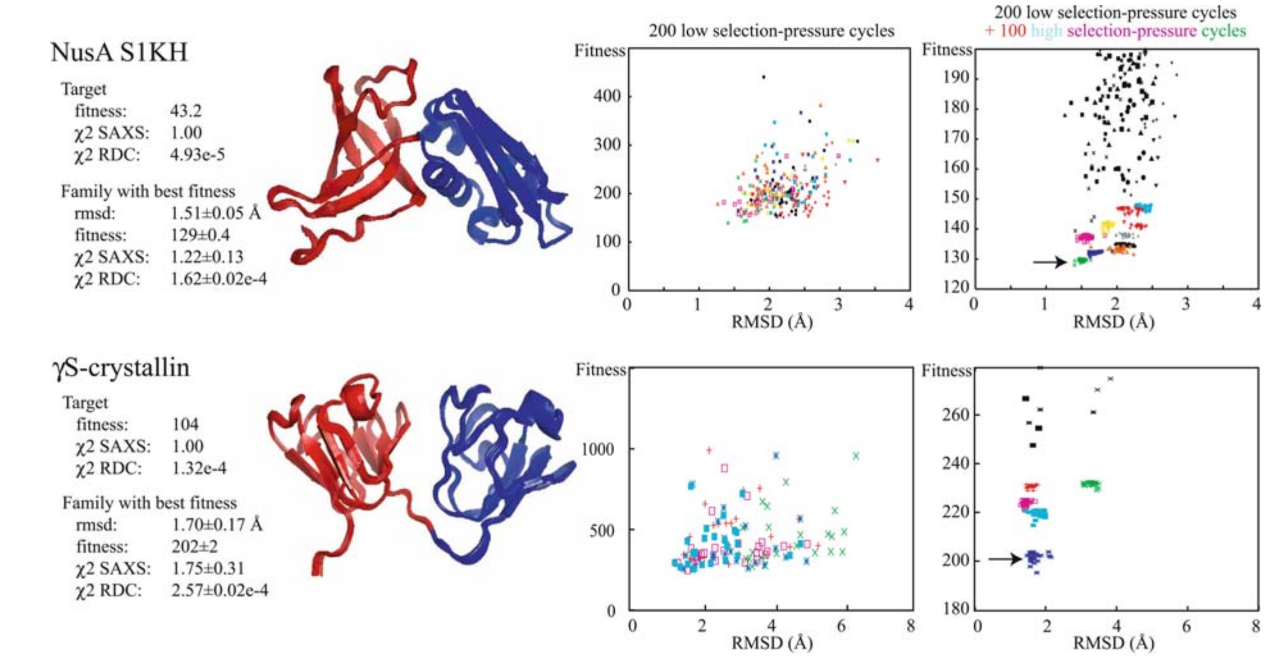


Fig. 3 Result of the genetic algorithm when starting from the target structures of *S1KH* and γS-crystallin. *Left side* for each protein, the family of structures with the lowest fitness is displayed. The rmsd to the target, the fitness and the separate contributions of *SAXS* and *RDC* are indicated. *Right side* we plotted the fitness versus the rmsd [calculated on all backbone atoms after superimposition of the conserved (B=1 and B=2) regions] at the end of the 200 first generations (*left plot/color symbols*; *right plot/black symbols*) and after 100 supplementary generations with high selection-pressure (*right plot/color symbols*). The displayed families are indicated by *arrows*

initial populations from the *T. maritima* S1 and KH crystallographic structures. The conformation of the last three amino-acids of S1 and the first three of KH was randomized and the domains were associated as described in the ‘‘Methods’’ section. We also examined if it was possible to discriminate between the four relative orientations compatible with RDC data. In fact, only two were possible, the others systematically leading to sterical clashes. Considering that in a real case there could be structural differences between isolated and associated domains, we treated them at deformable (i.e. we introduced mutations inside them).

The results obtained after the first 200 low pressure and following 100 high pressure cycles are presented in Fig. 4. We first observe that the genetic algorithm really acts as an optimizer and not as a simple filter. While the initial correctly oriented populations contain only one structure with an rmsd below 3 Å, most of the structures have rmsds of about 2 Å at the end of the low selection pressure cycles. Although rmsds are spread between 1.8 and 14 Å, the structures with the best fitnesses (160–200) are clustered around 2 Å. Similarly, at the end of the 100 additional high selective pressure cycles, all solutions have rmsd between 1.3 and 12 Å, but the three best families (fitness = 145, 148 and 151) have rmsds of 1.96, 2.41 and 2.31 Å. There is a family with a lower rmsd (1.30 Å), but it only ranks fifth in term of fitness (159), while the family in the fourth position (fitness = 156) has an rmsd of 3.80 Å.

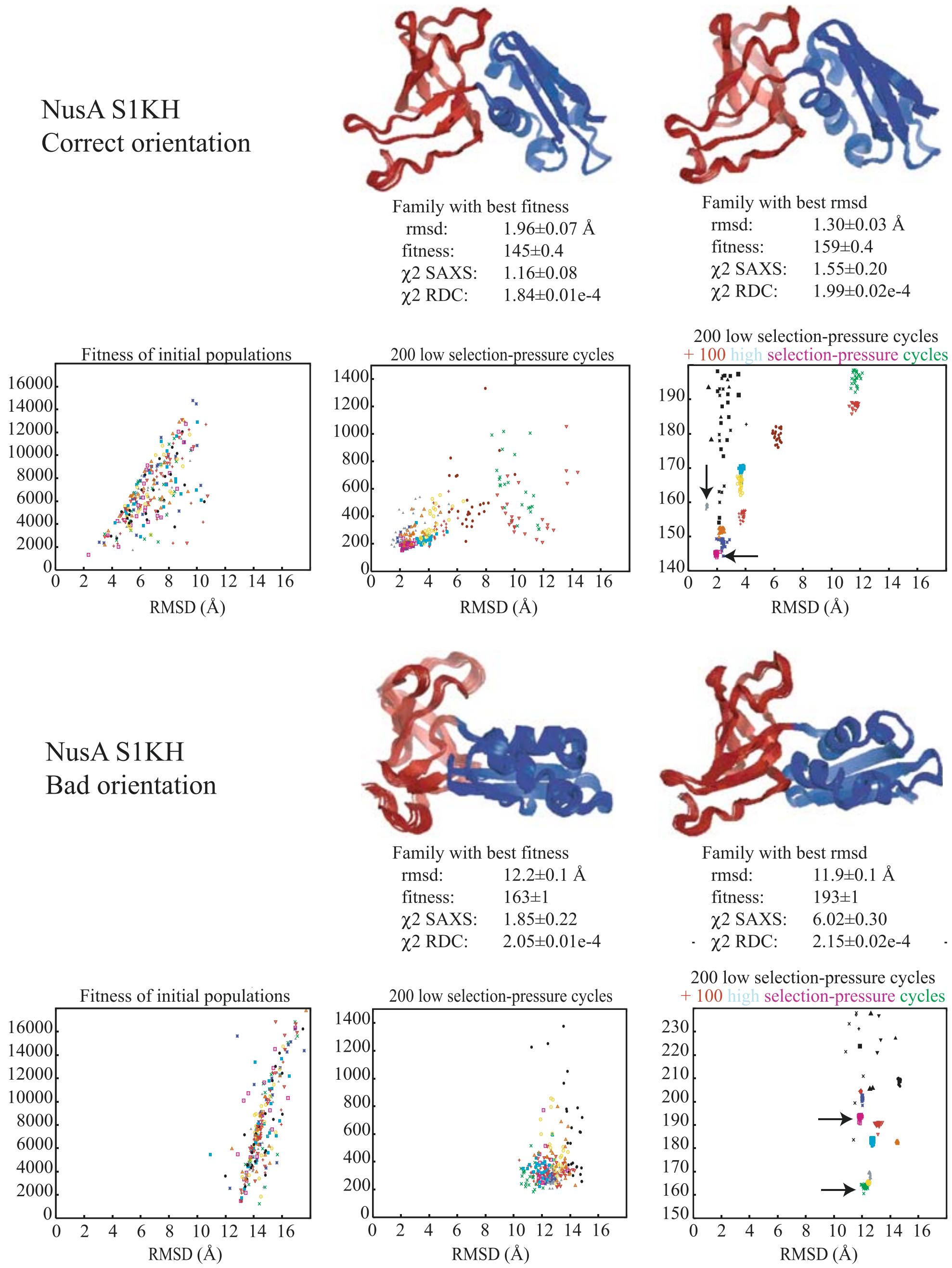
Finally, we remark that the best badly oriented family (in terms of fitness) has a higher fitness (163) than the three best correctly oriented. However, it is not sure that this result would be reproducible and that this can be really used to discriminate between the orientations.

Starting with individual domain homology models

Finally, we addressed the real situation by building the initial populations from homology models of the domains. For both proteins, we ran the genetic algorithm on ten correctly and ten wrongly oriented populations. As expected in the light of the previous results, we found that we were not able to discriminate between the different orientations arising from the degeneracy of RDCs. So, we only present the results obtained with the correct orientation in Fig. 5.

We first observe that the solutions have a higher fitness than in the previous essay (207 at best compared to 151 when starting from the individual X-ray modules and to 129 from the target structure of S1KH; 264 compared to 202 when starting from the target structure of γS-crystallin). Nevertheless, the procedure is able to generate solutions with low rmsds (better than in the initial populations).

Fig. 4 Results of the genetic algorithm when starting from the crystallographic structures of the domains. The crystallographic structures of the isolated *S1* and *KH* domains were used, after randomization of the linker regions, to form the initial populations as described in the text. *Upper part* the two domains are oriented like in the target structure. *Lower part* the KH domain is orientated with a 180° rotation around the *y*-axis. For each orientation, we displayed the families with the best fitness and with the best rmsd. We plotted the fitness versus the rmsd to the target in the initial population (*left plot*), at the end of the 200 first generations (*middle plot*/*color symbols*; *right plot*/*black symbols*) and after 100 supplementary generations with high selection-pressure (*right plot*/*color symbols*). The displayed families are indicated by *arrows*



At the end of the first 200 cycles, the rmsd spread of the low fitness solutions are 2–6 Å for S1KH and 3–7 Å for γS-crystallin. After the 100 supplementary high selective pressure cycles, the best three S1KH families (fitness = 208, 209 and 211) have rmsds of 2.67, 4.91 and 3.39 Å. In the case of γS-crystallin, the situation is more complicated. The two best families (fitness = 264 and 269) have rmsds of 4.28 and 5.60 Å, respectively, but the two following (both with a fitness = 272) have rmsds equal to 3.49 and 3.61 Å.

Discussion

Earlier multi-domain proteins have been assembled from the structures of individual domains by combining SAXS and RDC data. Annila et al. (Mattinen et al. 2002) have studied the structural modifications induced on calmodulin upon its binding by trifluoroperazin. Starting with the coordinates of the free form of the protein, they first oriented the N- and C-terminal domains with respect to the N-H^N RDCs by using an SVD procedure. They subsequently

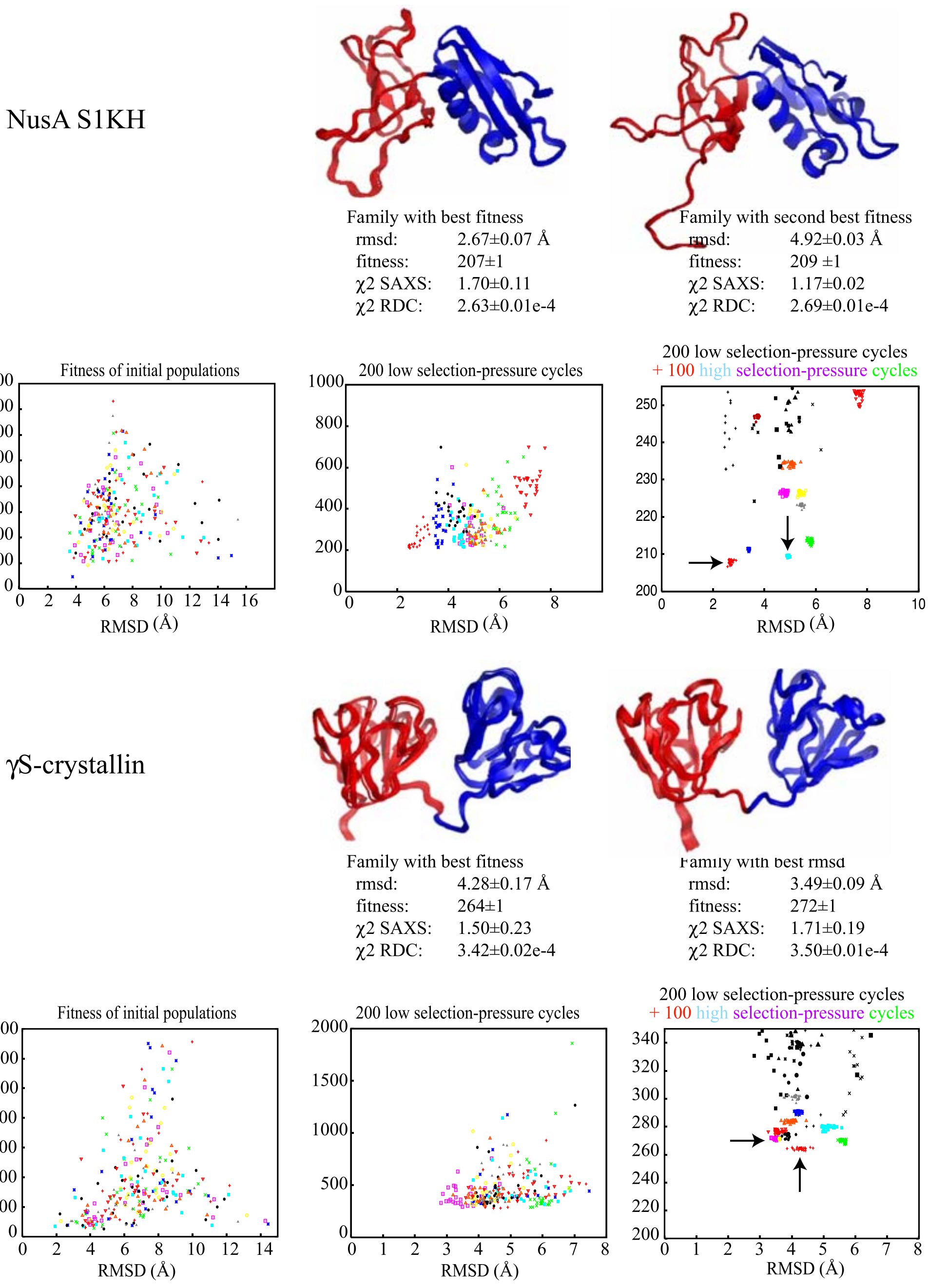


Fig. 5 Results from the homology models of the domains. Homology models of the domains were used to form the initial populations. *Upper part* results obtained on the *S1KH* bi-domain. *Lower part* results obtained on γS-crystallin. In both cases, the results correspond only to the initial populations with a relative orientation of the two domains similar to that observed in the target structures. We displayed the structures having the best and the second best fitness (in the case of *S1KH*) and the structures having the best fitness and the best rmsd (in the case of *γS-crystallin*). We plotted the fitness versus the rmsd to the target in the initial population (*left plot*), at the end of the 200 first generations (*middle plot*/*color symbols*; *right plot*/*black symbols*) and after 100 supplementary generations with high selection-pressure (*right plot*/*color symbols*). The displayed families are indicated by *arrows*

positioned the two domains with respect to the SAXS data using a grid search. Their best solution has an rmsd, calculated on the C$^{\alpha}$ atoms of residues 4–74 and 86–147, equal to 2.4 Å. Sattler et al. (Gabel et al. 2006) derived a parametric form of the beginning of SAXS curves (corresponding to three times the Guinier range) that can be incorporated as an energy term in structure calculation protocols. They applied their approach to the determination of the barnase/barstar complex structure from high-resolution structures of the free proteins, assuming that they could orient the proteins from RDC data, and using a simulated SAXS curve. They found two solutions, one corresponding to the target structure. But they also indicated that the identification of the solution depended on the crossing of two very similar curves, suggesting that in other cases, especially when docking spherical proteins, solution determination would be out of reach. Finally, Bax et al. (Grishaev et al. 2005) reported the structure determination of γS-crystallin by combining SAXS and NMR data. Using ‘‘globs’’ (i.e. small groups of atoms) and glob scattering factors, together with an iterative correction procedure allowing the minimization of the error introduced by their approximation, they were able to introduce a SAXS potential and its gradient in CNS. They used this to calculate the structure of γS–crystallin by simulated annealing in the presence of a small number of NOE distances (179 H^N-H^N, 70 CH_3-CH_3 including 15 inter-domain restraints), a set of RDCs recorded in two media (291 N-H^N, 303 C-C^α, 273 N-C’, 246 C^α-C^β) and a set of (ϕ,φ) dihedral angle restraints obtained from molecular fragment replacement (Kontaxis et al. 2005). In addition, their energy function contained a backbone–backbone hydrogen bonding potential (Grishaev and Bax 2004). They observe that the introduction of the SAXS potential induces a better compaction of the protein, leading to a better agreement between the NMR structure of γS-crystallin and the X-ray structures of γB-crystallin (rmsd calculated on the module backbone atoms of 1.31 Å in the presence of the SAXS data instead of 1.91 in their absence) and of γD-crystallin (rmsd of 1.18 Å instead of 1.89).

These approaches clearly demonstrate the power of combining SAXS and RDC data for the calculation of multi-domain protein structures. However, in all cases the authors either used a rigid representation of the known domain structures (Annila et al., Sattler et al.) or introduced additional data to define them (Bax et al.). We were interested to determine the structure of a bi-domain protein without knowing the exact structures of the domains (but with the assumption that homology models can be calculated) and by using only rather easily obtainable data. Recording of SAXS curves only requires the preparation of a concentration series of mono-disperse solutions to take into proper account interparticle interactions. Interpretation of RDCs only requires the assignment of backbone atom resonance frequencies. We restricted RDC measurement to a single medium at this stage and we did not consider the introduction of chemical shift mapping information, as this would require the production and purification of the independent modules. We also omitted information on the ϕ,φ angle that could be deduced from the backbone chemical shifts (Wishart and Case 2001), but it could be easily introduced, either during the homology modeling stage or during the optimization procedure.

We wanted to build a versatile tool, in which it would be easy to introduce any kind of additional information. Accordingly, we used a genetic algorithm as optimizer and chose a simple process: we limited the number of mutational operators to two (local and global backbone dihedral angle modifications), we built a fitness as a weighted sum of the individual evaluation functions and we chose a simple selection scheme. We also wanted to ensure that the procedure only generated plausible protein structures. After each mutation and recombination we minimized the covalent and van-der-Waals (but not the electrostatic) terms of the protein energy function and we introduced an energy cut-off to reject the structures that were too deformed. Finally, we also assumed that the quality of a homology model strongly varies along the sequence but that it is generally possible to discriminate between the well and less-well predicted regions. Accordingly, we limited the modifications in the well-predicted regions while we increased them in the others. We also took advantage of the information on the local quality of the structure given by the RDCs to modulate the mutation amplitudes.

The first result of our test is that despite its simplicity, our genetic algorithm is able to optimize the fitness of a bi-domain structure and to provide better solutions than those present in the initial population. However, the quality of the results clearly depends on the starting structures. They are better when starting from the X-ray structures of the whole protein (S1KH: best fitness = 129; γS-crystallin: best fitness = 202) or of the modules (S1KH: best fitness = 145) than when starting from the homology models (S1KH: best fitness = 207; γS-crystallin: best fitness = 264). In addition, we were never able to reach the values determined in the case of the target structures (S1KH: fitness = 43; γS-crystallin: fitness = 104). This suggests the existence of multiple minima in the evaluation function that strongly perturbs the search of the global minimum. But, by comparison with our first tests, we found that running the genetic algorithm in two steps, with a high selective pressure stage following a low selective pressure stage (by crude analogy with a simulated annealing protocol), greatly improved the convergence of the method and we expect that there is still room for improvement.

The second observation is that there is in all cases a net improvement of the structures. In the case of the X-ray domains of S1 and KH, all initial structures except one have rmsds between 3 and 10 Å. At the end of the process, four out of ten families contain structures with rmsds below or around 2 Å and seven out of ten families contain structures with rmsds below or around 4 Å. Similarly, all initial structures built from S1 and KH domain homology models have rmsds between 4 and 15 Å, while, at the end, three families contain structures with rmsds below 4 Å and nine rmsds below 6 Å. Finally, in the case of homology models of γS-crystallin domains, the initial populations contain structures with rmsds between 2 and 14 Å while all final structures have rmsds between 3 and 6 Å (in this case, the best structures of the initial populations were lost during the process). In addition, in all cases, the family with the best fitness has either the smallest value of rmsd (when starting from the S1KH X-ray structure, and S1KH homology models) or a value very close to the smallest (when starting from S1KH X-ray domains, γS-crystallin X-ray structure and γS-crystallin domain homology models). We have not obtained solutions corresponding to the target structure and we are not always able to identify the best solution by considering the fitness. However the final fitness value is probably too high but it can be improved by optimizing the selection pressures. It should be noticed that Bax et al. showed that the introduction of SAXS data leads to a precision of around 1 Å on their structure, but their structure was already well defined in the absence of SAXS. Their result is not contradictory with the idea that non-distinguishable minima could exist in the vicinity of the solution. In agreement with this idea, in the case of the study realized by Annila et al, the best identified solution has an rmsd calculated on the C^{α} atoms of the conserved region of calmodulin equal to 2.4 Å. This is in the same range than our result on S1KH when starting with the X-ray modules (2.0 Å calculated on all backbone of the whole protein) or with the homology models (2.67 Å).

In conclusion, we devised a versatile method based on a simple genetic algorithm to build bi-domain protein models from homologous structures on the basis of RDC and SAXS data. In the case of the S1KH bi-domain of the NusA protein, we identified a solution at 2.7 Å of the target by selecting the family having the best fitness. In the case of γS-crystallin, we miss the closest solution (3.5 Å) but obtain one at 4.3 Å. However, we used a very crude optimization scheme (a low selective pressure stage followed by a very high selective pressure stage) that could probably be improved. We are also not able to discriminate between the different relative orientations of the domains compatible with the RDC data. This can be solved by measuring RDCs in a second medium or by comparing the experimentally determined global alignment tensor with that obtained by prediction programs. We also did not try to take into account the dynamics of the system and assumed a single conformation, like in most studies. The point is that, by using data averaged over an ensemble of conformations, we do not expect to obtain a distribution of representative structures, but rather a ‘‘mean’’ solution. This could be possibly overcome by modifying the procedure and running the calculation with ‘‘individuals’’ formed of set of two structures. We will explore this possibility more deeply.

Aknowledgment We are grateful to Aldjia Mazari (Centre de Mathématiques Appliquées de l'Ecole Polytechnique) for helpful discussions. Fabien Mareuil was supported by a grant of the Institut de Chimie des Substances Naturelles (ICSN). Javier Perez benefited from the SAXIER program (Sixth Framework Program). The program suite is available upon request to the authors.

References

Bax A (2003) Weak alignment offers new NMR opportunities to study protein structure and dynamics. Protein Sci 12:1–16

Brünger AT (1992) X-PLOR: a system for X-ray crystallography and NMR. The Howard Hughes Medical Institute and Department of Molecular Biophysic and Biochemistry, Yale University

Eiben AE, Smith JE (2003) Introduction to evolutionary computing. Springer, Heidelberg

Ginalski K (2006) Comparative modeling for protein structure prediction. Curr opin struct biol 16:172–177

Gabel F, Simon B, Sattler M (2006) A target function for quaternary structural refinement from small angle scattering and NMR orientational restraints. Eur Biophys J 35:313–327

Grishaev A, Bax A (2004) An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. J Am Chem Soc 126:7281–7292

Grishaev A, Wu J, Trewhella J, Bax A (2005) Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. J Am Chem Soc 127:16621–16628

Kontaxis G, Delaglio F, Bax A (2005) Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining. Meth enzymol 394:42–78

Lipsitz RS, Tjandra N (2004) Residual dipolar couplings in NMR structure analysis. Annu rev biophys biomol struct 33:387–413

Losonczi JA, Andrec M, Fischer MW, Prestegard JH (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. J Magn Reson 138:334–342

Mattinen ML, Paakkonen K, Ikonen T, Craven J, Drakenberg T, Serimaa R, Waltho J, Annila A (2002) Quaternary structure built from subunits combining NMR and small-angle x-ray scattering data. Biophys J 83:1177–1183

Press WH, Teukolsky SA, Vatterling WT, Flannery BP (1992) Numerical recipes in C. 2nd edn. Cambridge University Press, Cambridge

Prestegard JH, Bougault CM, Kishore AI (2004) Residual dipolar couplings in structure determination of biomolecules. Chem Rev 104:3519–3540

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815

Svergun DI, Barberato C, Koch MHJ (1995) Crysol – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Cryst 28:768–773

Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science 278:1111–1114

Todd AE, Marsden RL, Thornton JM, Orengo CA (2005) Progress of structural genomics initiatives: an analysis of solved target structures. J Mol Biol 348:1235–1260

Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Meth Enzymol 338:3–34