Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A genetic algorithm for the identification of conformationally invariant regions in protein molecules

Thomas R. Schneider

Department of Structural Chemistry, University of Göttingen, Tammannstrasse 4, 37077 Göttingen, Germany

Correspondence e-mail: trs@shelx.uni-ac.gwdg.de

Received 27 July 2001
Accepted 12 November 2001

Understanding macromolecular function often relies on the comparison of different structural models of a molecule. In such a comparative analysis, the identification of the part of the molecule that is *conformationally invariant* with respect to a set of conformers is a critical step, as the corresponding subset of atoms constitutes the reference for subsequent analysis for example by least-squares superposition. A method is presented that categorizes atoms in a molecule as either conformationally invariant or flexible by automatic analysis of an ensemble of conformers (*e.g.* crystal structures from different crystal forms or molecules related by non-crystallographic symmetry). Different levels of coordinate precision, both for different models and for individual atoms, are taken explicitly into account *via* a modified form of Cruickshank's DPI [Cruickshank (1999), *Acta Cryst.* D**55**, 583–601] and are propagated into error-scaled difference distance matrices [Schneider (2000), *Acta Cryst.* D**56**, 715–721]. All pairwise error-scaled difference distance matrices are then analysed simultaneously using a genetic algorithm. The algorithm has been tested on several well known examples and has been found to converge rapidly to reasonable results using a standard set of parameters. In addition to the description of the algorithm, a criterion is suggested for testing the identity of two three-dimensional models within experimental error without any explicit superposition.

1. Introduction

Although a single crystal structure normally represents a rather static picture of a protein molecule, the comparison of different conformers of a given molecule as obtained from different structure determinations can provide valuable insight into its flexibility. Such analyses can be based on structures in different crystal forms, under different physicochemical conditions (using, for example, pH or temperature as a variable) and with and without a substrate molecule bound. Multiple copies of a molecule may also arise from non-crystallographic symmetry or be an ensemble generated by NMR structure analysis. A typical strategy is to first identify the rigid part of the molecule and then in a subsequent step interpret the conformational differences revealed by least-squares superposition of the corresponding atoms in the rigid part. There are, however, several problems with this approach. Firstly, the identification of the rigid or *conformationally invariant* part of the molecule, here defined as the largest subset of atoms for which all interatomic distances are identical across all models under consideration, is often performed manually and in an iterative fashion and is thus susceptible to preconceived ideas. Secondly, in most cases, the potentially

very different levels of quality of different models and of the precision of atomic coordinates for different atoms in the same model are not taken into account in the comparison process. A third, more technical, complication is that usually the different models are compared in a pairwise fashion, resulting in massive book-keeping problems if many models are compared. Recently, an approach to identify the conformationally invariant part of a molecule with respect to a set of several conformers by manual interpretation of error-scaled difference distance matrices has been proposed (Schneider, 2000). In principle, this approach can be used to identify the conformationally invariant part of a molecule with respect to any number of conformers; in practice, however, the number and the complexity of the matrices to be interpreted soon become the limiting factor. In the present paper, a computer algorithm that automizes this task is described.

The question of how to extract information about the rigidity and flexibility of proteins based on the comparison of *two* conformers has already been addressed by several authors. The *DYNDOM* approach developed by Berendsen & Hayward (1998) employs a cluster analysis of interatomic vectors to first subdivide a molecule into domains and then determines hinge axes between these domains. Lesk has described the *sieve fitting procedure* (Lesk & Chothia, 1984; McPhalen, Vincent & Jansonius, 1992), where the dominant rigid domain of a molecule is found by iterative application of least-squares fits in which after starting from a full superposition in every step poorly matching residues are eliminated. This method has been expanded to the partitioning of a molecule into multiple small domains of preserved geometry in the *adaptive selection* procedure suggested by Wriggers & Schulten (1997). Here, variable seed sets of atoms are subjected to sieve fitting in order to develop a set of well fitting substructures. Nichols *et al.* (1995) have recently formulated two algorithms that find rigid domains by analysing the *thresholded difference distance matrix* corresponding to two conformers and applied this technique to identify rigid regions in haemoglobin based on the comparison of the respective deoxy and oxy structures (Nichols *et al.*, 1997).

A generalization of the sieve fitting procedure to the case of *multiple* prealigned models of different but homologous proteins has been put forward by Gerstein & Altman (1995) and has subsequently been used to create a library of protein family core structures (Schmidt *et al.*, 1997).

Multiple models of the same molecule also appear in ensembles of conformers derived from an NMR experiment. Kelley and coworkers have recently designed an algorithm for the analysis of such ensembles that finds the core of a molecule by identifying dihedral angles showing small variation across the ensemble. However, as in their definition of the core the parts of the molecule containing stable dihedral angles do not need to be continously connected, such a core can contain rigid regions that move relative to one another. Kelley and coworkers deal with such situations by further subdividing the core into local structural domains by analysis of a matrix containing the variance of $C^{\alpha}-C^{\alpha}$ distances. This matrix effectively summarizes the information present in the difference distance matrices between all possible pairs of conformers and can be rapidly interpreted using a clustering algorithm (Kelley *et al.*, 1996, 1997).

In the context of error-scaled difference distance matrices, finding the largest conformationally invariant part of a molecule with respect to a set of conformers (as opposed to only two conformers) can be formulated as an optimization problem of identifying the largest subset of atoms for which all interatomic distances are identical within error in all models (corresponding to all relevant elements of the error-scaled difference distance matrices being approximately zero). As for a molecule with N atoms the number of possibilities for creating such a subset is 2^N (every atom can be a member of the subset or not), a systematic full search of all possibilities is clearly not feasible. Furthermore, the fact that the search parameters assume discrete values prohibits an analytical solution of the problem. The nature of the problem, however, is such that local partial solutions (*i.e.* small groups of conformationally invariant atoms) can be identified and their combination may yield a better solution (*i.e.* if two rigid groups do not move relative to one another, their combination will result in a larger rigid group), making a genetic algorithm (GA; Holland, 1975; Goldberg, 1989; Mitchell, 1996) the optimization method of choice. In addition, GAs are also particularly well suited to large search spaces with discrete parameters.

The basic idea of a genetic algorithm is that a population of candidate solutions, or hypotheses, is created and then subjected to an evolutionary process. During the evolution, offspring candidate solutions are generated by combining properties of existing candidate solutions whereby the probability of properties being passed on to the next generation is proportional to a measure of fitness. After a number of generations, the evolutionary search should converge to a homogeneous population with high values of the fitness function for all members of the population. Although there is no guarantee that the global optimum is found, the top-scoring hypothesis of the final population in many cases represents an acceptable solution to the optimization problem.

A number of problems in structural biology have been tackled using GAs in recent years [*e.g.* protein form reconstruction from X-ray solution scattering (Chacón *et al.*, 2000) and docking of small molecules to macromolecules (Jones *et al.*, 1997)] and several applications in macromolecular crystallography have been reported [*e.g.* finding heavy-atom sites from Patterson maps (Chang & Lewis, 1994), molecular replacement (Chang & Lewis, 1997; Kissinger *et al.*, 1999) and low-resolution phasing (Webster & Hilgenfeld, 2001)].

The implementation of the search for the largest conformationally invariant part of a molecule as a GA is relatively straightforward. A candidate solution can be conveniently described as a binary string where every bit indicates whether or not the corresponding atom belongs to the conformationally invariant part or not. The fitness of a candidate solution is evaluated by checking its consistency with all error-scaled difference distance matrices that can be constructed between all conformers being analysed. The main difficulty in manual

interpretation, the combination of small conformationally invariant regions, where all atoms are close in sequence, to larger parts connected in three dimensions [while keeping track of a large number of error-scaled difference distance (EDD) matrices], is overcome by the genetic algorithm which, in fact, is centred exactly around the filtering of improved candidate solutions from a large pool of randomly recombined potential partial solutions.

In the following, the algorithm is described and its application to several test cases, for which a careful more manual analysis was available in the literature, is discussed. The examples are (i) comparison of three NCS-related copies of chorismate mutase refined to near atomic resolution, (ii) comparison of two NCS-related copies of an Fab fragment, (iii) determination of the large and small domain of aspartate aminotransferase based on five structures representing different ligation states in four different crystal forms, (iv) derivation of the common core of seven different structures of pig pancreatic α-amylase in four different crystal forms and (v) derivation of rigid and flexible parts of the enzyme epimerase based on ten conformers related by non-crystallographic symmetry.

In all examples, the method is used for comparing C^{α} atoms only. It should be noted that the method is applicable to any set of atoms, such as, for example, atoms surrounding the active site of an enzyme.

2. Methods

2.1. Formulation of the optimzation problem

The standard difference distance matrix contains elements Δ_{ij}^{ab} corresponding to the difference in distance between two atoms i and j in two conformers $\mathcal{M}_a$ and $\mathcal{M}_b$,

$$\Delta_{ij}^{ab} = d_{ij}^{a} - d_{ij}^{b}. \tag{1}$$

If estimates $\sigma(\Delta_{ij}^{ab})$ for the uncertainty of the matrix element Δ_{ij}^{ab} are available, the corresponding error-scaled difference distance with elements E_{ij}^{ab} can be calculated (Schneider, 2000),

$$E_{ij}^{ab} = \Delta_{ij}^{ab}/\sigma(\Delta_{ij}^{ab}). \tag{2}$$

A rough approximation for $\sigma(\Delta_{ij}^{ab})$ has been suggested in Schneider (2000) based on the coordinate uncertainty of atoms i and j in models $\mathcal{M}_a$ and $\mathcal{M}_b$, $\sigma_{x,i}^{a}$, $\sigma_{x,j}^{a}$, $\sigma_{x,i}^{b}$ and $\sigma_{x,j}^{b}$,

$$\sigma(\Delta_{ij}^{ab}) = \left[(\sigma_{x,i}^{a})^2 + (\sigma_{x,j}^{a})^2 + (\sigma_{x,i}^{b})^2 + (\sigma_{x,i}^{b})^2\right]^{1/2}. \tag{3}$$

Values for atomic coordinate uncertainties $\sigma_{r,i}^{a}$ can be calculated rigorously *via* the inversion of the normal matrix of the refinement (Sheldrick & Schneider, 1997) or estimated using a modified version of Cruickshank's Diffraction Precision Indicator, DPI (Cruickshank, 1999; Schneider, 2000). According to Cruickshank, the coordinate error of an atom with a mean B value $B_{\rm avg}$, $\sigma(x, B_{\rm avg})$, can be estimated based on the number of fully occupied sites N_i, the number of unique reflections $n_{\rm obs}$, the completeness of the data C, the free R value $R_{\rm free}$ and the maximum resolution of the data $d_{\rm min}$ (equation 27 in Cruickshank, 1999),

$$\mathrm{DPI}_f = \sigma_f(x, B_{\rm avg}) = (N_i/n_{\rm obs})^{1/2} C^{-1/3} R_{\rm free} d_{\rm min}. \tag{4}$$

In cases where $R_{\rm free}$ is not known and the number of parameters $n_{\rm par}$ is smaller than the number of observables $n_{\rm obs}$, equation (26) from Cruickshank (1999),

$$\mathrm{DPI}_a = \sigma_a(x, B_{\rm avg}) = (N_i/p)^{1/2} C^{-1/3} R d_{\rm min}, \tag{5}$$

where $p = (n_{\rm obs} - n_{\rm par})$ and R equals the crystallographic R value, can be used to derive an estimated error. In a first approximation, assuming that coordinate uncertainties and B values are linearly related, the coordinate uncertainty for an individual atom with a B value of B_i can be estimated as

$$\tilde{\sigma}_{af,i} = \frac{\mathrm{DPI}_{af}}{B_{\rm avg}} B_i, \tag{6}$$

depending on whether DPI_a or DPI_f has been evaluated. To avoid extremely low errors for atoms with extremely low B values (that may well correspond to refinement artefacts), B values smaller than a certain value $B_{\rm low}$ can be scaled up to $B_{\rm low}$. In the computer program implementing the algorithm, the default value of $B_{\rm low}$ is set to $B_{\rm avg} - 2\sigma_B$, where σ_B is the standard deviation of $B_{\rm avg}$.

The goal of finding the conformationally invariant part of a molecule with respect to a set of n structural models $\mathcal{M} = \{\mathcal{M}_1 \ \ldots \ \mathcal{M}_n\}$ is equivalent to identifying a subset of the set of all atoms $\mathcal{A} = \{a_1 \ \ldots \ a_{n_{\rm atom}}\}$ for which for all pairs of atoms ij the interatomic distance are identical in all models. Allowing for distances being imprecise owing to experimental errors, this condition can be relaxed to the condition that the difference in distances should not be significantly different from zero; in other words, for all pairs of models ab and all pairs of atoms ij belonging to the conformationally invariant subset, the respective elements of the error-scaled difference distance matrices E_{ij}^{ab} should be smaller than a tolerance level ε_l.

2.2. Implementation as a genetic algorithm

2.2.1. Encoding, fitness function and population. A candidate solution, or hypothesis, $\mathcal{H}$ about a subset of atoms being conformationally invariant can be conveniently encoded as a string $\mathcal{H}$ of length $n_{\rm atom}$ over the binary alphabet {0, 1},

$$\mathcal{H} = \{h_1 \ldots h_{n_{\rm atom}}\} \tag{7}$$

where the value or status of a bit h_i depends on whether or not (0 or 1) an atom i is considered to be a member of the conformationally invariant subset.

Based on the error-scaled distance differences, E_{ij}^{ab}, the fitness of an hypothesis can be measured using the following fitness function $S_{\mathcal{H}}$,

$$S_{\mathcal{H}} = \frac{1}{n_{ab} n_{\rm atom}^2 \varepsilon_l} \sum_{ab} \sum_{ij} \begin{cases} \varepsilon_l - |E_{ij}^{ab}| & \text{for } |E_{ij}^{ab}| < \varepsilon_l \\ p \times \min(|E_{ij}^{ab}|, \varepsilon_h) & \text{for } |E_{ij}^{ab}| \geq \varepsilon_l \end{cases}. \tag{8}$$

This score is a weighted count over all relevant elements of all error-scaled difference matrices (the outer sum running over all possible pairs *ab* of models $\mathcal{M}_a$ and $\mathcal{M}_b$, the inner sum running over all elements *ij* of the respective EDD matrices that should be zero according to the hypothesis $\mathcal{H}$). For every matrix element that is consistent with the hypothesis, *i.e.* $|E_{ij}^{ab}| < \varepsilon_l$, the score is increased by a term $(\varepsilon_l - E_{ij}^{ab})$. For matrix elements not consistent with the hypothesis, *i.e.* $|E_{ij}^{ab}| \geq \varepsilon_l$, the score is reduced by the size of the respective element scaled by a penalty factor p. The choice of the penalty factor (typically between 5 and 20) allows one to adjust the fraction of non-zero matrix elements allowed for an acceptable solution. The matrix elements entering the penalty term are limited to a maximum of ε_h in order to avoid instabilities in the evolutionary search caused by inclusion of single well defined atoms with extremely large conformational differences. Without this mechanism, inclusion of a single such atom into a hypothesis would have an extreme effect on the score, possibly rendering an otherwise well performing candidate solution useless. The normalization factor $1/(n_{ab}n_{\rm atom}^2\varepsilon_l)$ is chosen such that for the case where the entire molecule is rigid and all elements of all error-scaled difference matrices are zero, a score of 1.0 is obtained.

Throughout the evolutionary search, a population of $n_{\rm hyp}$ hypotheses (where a typical value of $n_{\rm hyp}$ is 20) $\mathcal{H}_k$ is maintained.

2.2.2. Genetic operators. To modify the population, the following genetic operators are used.

(i) *Grow/Shrink*. Moving from left to right, each bit of an $\mathcal{H}$-string is inverted if it corresponds to the beginning or end of a stretch of atoms continous in sequence marked as conformationally invariant. The corresponding bit is kept inverted only if it results in an increased fitness. This operation is repeated iteratively until no further improvement of the score is achieved.

(ii) *Recombination*. Two hypotheses are taken, subdivided into ten non-overlapping regions (genes) of equal length and an offspring is constructed by recombining genes taken randomly from the two parent hypotheses.

(iii) *Mutation*. The status of a randomly chosen atom is inverted. The mutation is only kept if it leads to an improvement in the score $S_{\mathcal{H}}$. In principle, in an evolutionary process mutations that lead to a temporary deterioration of the score may be beneficial in the long run, but experience has shown that for the present purpose permitting disadvantageous mutations can sometimes render the optimization process unstable (see §3.1). Permitting only advantageous mutations also speeds up convergence. A typical mutation rate is 5%.

2.2.3. The starting population. A starting population of $n_{\rm hyp}$ hypotheses is created by marking continous overlapping stretches of length $1.2 \times n_{\rm atom}/n_{\rm hyp}$ atoms as conformationally invariant. The Grow/Shrink operator is then applied to all hypotheses of the starting population and the central loop of the evolutionary algorithm implementing selection and reproduction is entered.

2.2.4. Evolution. A new generation is created from an existing population, modified and evaluated in six steps.

(i) *Selection*. After sorting the hypotheses with respect to their score, the low-scoring 30% of the hypotheses are destroyed and only the top-scoring 70% of the hypotheses are kept for subsequent reproduction.

(ii) *Recombination including crossover*. The now empty 30% of the slots are replenished by random recombination from the surviving hypotheses using the Recombination operator. To assign a higher mating probability for high-scoring individuals, each parent is selected by taking the highest scoring hypotheses of two successive random draws from the survivors of the previous generation.

(iii) *Scoring*. All $n_{\rm hyp}$ hypotheses are scored.

(iv) *Mutation*. All hypotheses undergo random mutagenesis where only mutations improving the fitness are kept (see Mutation operator above).

(v) *Grow/Shrink*. All members of the population develop using the Grow/Shrink operator.

(vi) *Convergence check*. Several criteria to stop the optimization process are evaluated. If no convergence criterion is fulfilled, the next generation is started from (i) (Selection).

2.2.5. Convergence. In most cases, the optimization will be terminated when the population is becoming homogeneous, *i.e.* most hypotheses are similar. Although identical scores are not necessarily equivalent to identical hypotheses for the general case, in the case of the present algorithm the homogeneity of a population can be measured as the standard deviation of its scores. If this standard deviation is smaller than a user-defined percentage $p_{\rm std}$ (by default, $p_{\rm std} = 1.0\%$) of the mean score, further cycles are not likely to add any new information to the top-scoring hypothesis and the iteration is stopped.

To avoid the optimization becoming trapped in an oscillating situation, particularly in cases where very noisy matrices are analysed, the course of the top score is monitored. If the top score does not change for more than a number of generations (typically ten), the search is stopped.

Finally, if the number of generations reaches a user-defined limit (*e.g.* 50 generations), the search is interrupted.

2.2.6. Constraints. Acceptable solutions are not only characterized by a high score, but also should fulfil certain additional conditions to be physically reasonable. Two such conditions can be formulated: (i) conformationally invariant atoms should appear in consecutive stretches of a certain minimum length $\mathrm{sl}_{\rm min}$ and (ii) continous stretches of conformationally invariant atoms should not be interrupted by single-atom outliers.

As the overall performance of the algorithm largely depends on the speed of the scoring, no constraints that would complicate the evaluation of the fitness function were implemented in the evolutionary search itself. Constraints are only imposed after the search has finished by modification of the top-scoring hypothesis: stretches shorter then $\mathrm{sl}_{\rm min}$ atoms are marked as flexible and the status of single flexible atoms in otherwise sequence-continous stretches of conformationally invariant atoms are inverted. These constraints are of course only valid if the order of the atoms analysed has a physical meaning, as for example is normally the case for C^{α} atoms,

which belong to successive amino-acid residues in a polypeptide chain; for cases where the sequence of atoms analysed is not related to their arrangement in three dimensions, these constraints should not be applied.

2.3. Identification of identical models

An important question that can be answered in terms of error-scaled difference is whether two models as a whole can be considered identical or not. Again, in principle, for two identical models all elements of the EDD matrix should be zero. In reality, this condition can be relaxed twofold to allow for experimental uncertainties: (i) elements whose modulus is smaller than a threshold ε_l are considered to be zero within error and (ii) a certain percentage of elements of the EDD matrix is allowed to have values larger than ε_l.

If we consider two models consisting of N atoms each, where M of the N atoms are in significantly different positions, the number of zero-elements n_0 in the corresponding EDD matrix should still be larger than

$$n_0 = (N - M)^2 = N^2 - 2MN + M^2 \qquad (9)$$

[where $(N - M)^2$ is the number of elements in the block of the matrix that corresponds to the conformationally invariant part of the model], or, expressed as a percentage of zero elements with respect to all elements,

$$p_0 = \frac{N^2 - 2MN + M^2}{N^2} = 1 - 2\frac{M}{N} + \left(\frac{M}{N}\right)^2. \qquad (10)$$

If we now replace M/N by the percentage of atoms that are significantly different, $p_d = M/N$, we obtain the following equation for the corresponding percentage of EDD-matrix elements that should still be zero if a percentage p_d of atoms is in different conformations in the two conformers being compared,

$$p_{0,p_d} = 1 - 2p_d + p_d^2. \qquad (11)$$

If we allow 1% of the atoms to be in different relative positions (*i.e.* $p_d = 0.01$), we obtain

$$p_{0,1\%} = 1 - (2 \times 0.01) + 0.01^2 = 0.9801. \qquad (12)$$

2.4. Computers and other programs used

All calculations were performed on a PC with a Pentium III processor running at 800 MHz under Linux Kernel Version 2.2.16. All least-squares superpositions were performed using *LSQKAB* (Collaborative Computational Project, Number 4, 1994; Kabsch, 1976). Schematic figures of protein molecules were prepared using *MOLSCRIPT* (Kraulis, 1991) and *Raster3D* (Merritt & Murphy, 1994).

3. Examples

3.1. Chorismate mutase

The structure of the functional unit of chorismate mutase from *Bacillus subtilis*, a homotrimer, has recently been determined to a resolution of 1.3 Å (Ladner *et al.*, 2000). The crystals used belonged to space group $P2_12_12_1$ (unit-cell parameters a = 52.2, b = 83.8, c = 86.0 Å) with three crystallographically independent monomers in the asymmetric unit. Out of the 127 residues in a monomer, the first 116 residues form a compact structure of β-strands and helices; the C-terminal residues 117–127 are found in different conformations in each of the three monomers (Fig. 2).

Table 1
Data used for the calculation of Cruickshank's DPI for the structures investigated in this paper.

Values for the number of fully occupied site N_i, the number of observables n_{obs}, the completeness of the diffraction data cpl., the crystallographic R value for all data R, the free R value R_{free} and the maximum resolution d_{min} were taken from the headers of the respective PDB files unless otherwise indicated. n_{par} was calculated as $4 \times N_i$ (assuming refinement of three coordinates and one B value per atomic site) for all structures except 1dbf. DPI_f and DPI_a were calculated using (4) and (5), respectively.

PDB code	N_i	n_{par}	n_{obs}	cpl. (%)	R (%)	R_{free} (%)	d_{min} (Å)	DPI_a (Å)	DPI_f (Å)
1dbf†	3658	32922	89868	92.0	16.9	23.5	1.3	0.057	0.063
8fab	7073	28292	64477	75.0	17.3	n/a	1.8	0.15	n/a
7aat	6992	27968	60850	96.1	16.6	n/a	1.9	0.15	n/a
1tar‡	6352	25408	36893	88.5	19.4	n/a	2.2	0.35	n/a
1ama	3473	13892	17538	94.4	15.9	n/a	2.3	0.36	n/a
1tas‡	6096	24384	17636	87.9	16.0	(30.0)	2.8	n/a	(0.53)
1tat‡	6098	24392	18194	97.0	15.0	(30.0)	3.0	n/a	(0.54)
1pig	4257	17028	44944	93.7	17.8	21.0	2.2	0.16	0.15
1pif	4184	16736	40433	94.7	17.1	20.8	2.3	0.17	0.16
1ose	4377	17508	37212	90.0	17.6	22.2	2.3	0.20	0.18
1jfh	4363	17453	33421	99.4	16.0	18.5	2.03	0.17	0.14
1ppi	4358	17432	25018	95.0	15.3	n/a	2.2	0.26	n/a
1dhk§	5840	23360	66557	96.7	18.3	22.0	1.85	0.13	0.12
1bvn	4606	18424	17259	73.7	16.6	26.0	2.5	n/a	0.372
1eq2	25435	101740	242238	93.3	21.2	26.2	2.0	0.19	0.17

† As the model was refined with anisotropic displacement parameters for all non-H atoms, n_{par} was set to $9 \times N_i$ and cpl. was taken from the publication. ‡ All values were taken from the publication (Hohenester & Jansonius, 1994). § Value for cpl. was taken from the publication.

DPI values were calculated based both on R and R_{free} (Table 1) and the resulting R_{free}-based value was found to be somewhat higher than the value obtained using R. As DPI_f is more robust with respect to overfitting (which in fact could be the case to a small extent for a full anisotropic model with data extending to 1.3 Å) than DPI_a, all subsequent calculations are based on the DPI_f.

For the three molecules, three error-scaled $C^\alpha C^\alpha$ difference distance matrices with 8001 independent elements each were calculated and analysed. The uncertainties for the differences in $C^\alpha C^\alpha$ distances $\sigma(\Delta_{ij}^{ab})$ (3) ranged from 0.048 Å for atom pairs with low B factors to 0.285 Å for atom pairs with high B factors, with a mean value of 0.097 Å. The corresponding error-scaled difference distance matrix cut at a lower limit of ε_l of 5σ (Fig. 1*a*) very clearly shows that the major difference between the molecules A and B is the conformation of about ten residues at the C-terminus. Beside this, the second half of the first helix (residues 28–33) and a short loop preceding the C-terminus (residues 101–104) are marked as having different conformations relative to the major part of the protein.

Using standard parameters ($n_{hyp} = 20$, $w_p = 20.0$, $r_{mut} = 5.0\%$) and lower and upper tolerance levels of $\varepsilon_l = 5.0$ and $\varepsilon_h = 10.0$, respectively, the GA identified (using 5.0 s of CPU time), in very good agreement with the manual interpretation of the EDD matrix, 103 of the 127 atoms as conformationally invariant (Fig. 1*a*). All these residues (2–27, 34–101, 105–113) are located in the N-terminal domain of the protein (Fig. 2).

When the limits for display and the automatic search were reduced to $\varepsilon_l = 2.0$ and $\varepsilon_h = 5.0$, many more matrix elements showed a signal and manual interpretation became substantially more complicated if not impossible (Fig. 1*b*). Running the automatic procedure using the smaller tolerance levels and standard parameters otherwise ($n_{hyp} = 20$, $w_p = 20.0$, $r_{mut} = 5.0\%$), the rigid part was found to be substantially smaller. Residues 3–13, 48–52 and 86–95, *i.e.* only 26 residues, exclusively located in a small region of the central β-sheet, were now identified. This subset of residues is not completely continous in sequence but is nevertheless continous in three-dimensional space (Fig. 2). Here it should be noted that apart from the application of the constraints to atoms close in primary sequence as described in §2.2.6, no information about connectivity in three dimensions is used in the algorithm.

The C^{α} atoms of these 26 residues are sufficient to afford a robust superposition using standard least-squares superposition techniques, yielding mean r.m.s. deviations of 0.10 and 0.11 Å for superimposing molecule *B* and molecule *C* onto molecule *A*, respectively. The overall picture of a floppy C-terminus stays valid, while the superposition based on this small subset reveals small but significant conformational differences in the at first sight rigid N-terminal domain. This observation relates to the fact that at a resolution of 1.3 Å, at least for atoms with low *B* factors, changes in relative position of the order of 0.1 Å are definitely significant.

In this example, the differences in interatomic distances are rather large for some atoms in the C-terminal region (up to more than 10 Å corresponding to more than 100σ of the interatomic distance in some cases). In an early version of the algorithm, inclusion of an atom from this region often caused the evolutionary search to become unstable because well performing hypotheses were ruined by inclusion of one 'wrong' atom during the phase of random mutation. This

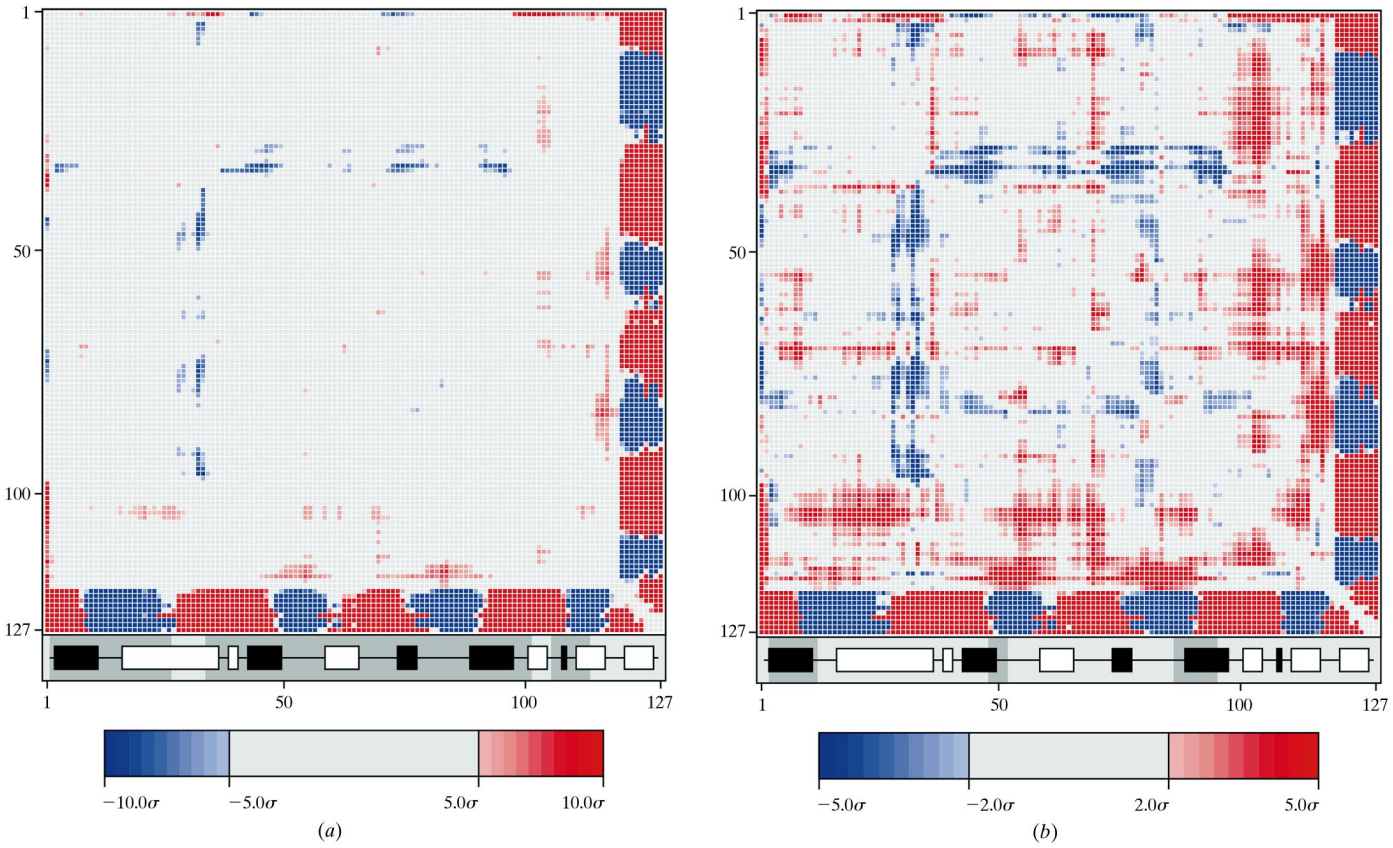

Figure 1
Error-scaled difference distance matrices for molecules *A* and *B* of chorismate mutase. All changes in distances smaller than a threshold (5σ and 2σ for *a* and *b*, respectively) are shown in grey; differences between this lower limit and an upper limit of 10σ (5σ for *b*) are shown using a colour gradient where red stands for expansion and blue for contraction, light colours represent small changes and dark colours large changes; all differences larger than the upper limit are shown as full blue and full red, respectively. The gradients used for colour coding are also shown separately at the bottom of the figure. The bar underneath the matrix in the foreground shows the secondary structure as white (α-helices) and black (β-strands) rectangles. The background of the secondary-structure scheme indicates parts of the protein that were identified to be conformationally invariant (dark grey) using a tolerance ε_l of 5σ (*a*) or 2σ (*b*), respectively. For clarity, the matrices underwent 2 × 2 binning (maintaining the element with the highest absolute value in the respective binning area) before being displayed.

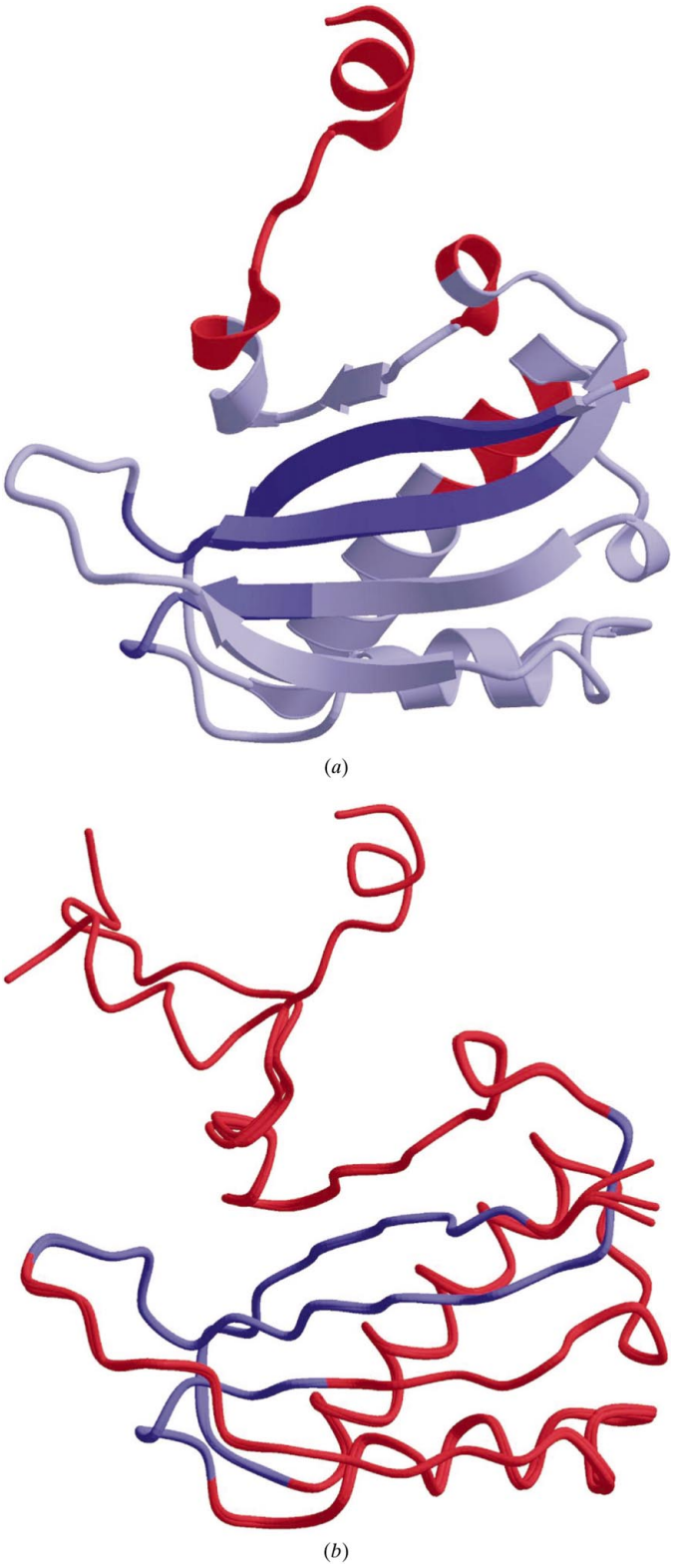

Figure 2
(*a*) Secondary structure of chorismate mutase. The colours correspond to the flexibility of the different parts of the molecule: dark blue shows regions that are conformationally invariant when a tolerance of $\varepsilon_l = 2\sigma$ is used, light blue marks parts that are conformationally invariant at the 5σ level and flexible regions are shown in red. (*b*) Superposition of the three crystallographically independent molecules of chorismate mutase using the 26 C^{α} atoms (indicated as a blue backbone trace) that were found to be conformationally invariant at the 2σ level.

problem was overcome by limiting the penalty term (8) and by only accepting mutations that improve the score of a hypothesis.

3.2. Fab fragment

Antibodies consist of immunoglobulin domains connected by hinge regions allowing variability in the orientation of the domains with respect to each other. As, owing to this flexibility, the crystallization of intact antibodies is rather difficult, many studies of antibodies are based on the crystal structure of Fab fragments (Branden & Tooze, 1999). In the case of an Fab fragment from human myeloma immunoglobulin studied by Saul & Poljak (1992), orthorhombic crystals in space group $P2_12_12_1$ (unit-cell parameters a = 110.6, b = 127.4, c = 66.5 Å) contain two Fab fragments in the asymmetric unit.

The DPI value based on the R value against all data was calculated to be 0.15 Å (Table 1), resulting in estimates for the coordinate errors for C^{α} atoms ranging between 0.048 and 0.389 Å.

The error-scaled difference distance matrix for the comparison of the two Fab fragments (Fig. 3*a*) clearly reveals that the Fab fragment consists of four mostly rigid domains that can move relative to one another: the matrix contains four 'empty' blocks along the diagonal corresponding to the variable (N-terminal) and the constant (C-terminal) parts of both the light (chain *A*) and the heavy (chain *B*) chain of the antibody.

Using standard parameters ($n_{\rm hyp} = 20$, $w_p = 20.0$, $r_{\rm mut} = 5.0\%$) and lower and upper tolerance levels of ε_l = 2.0 and ε_h = 5.0, respectively, the GA identified (using 2.0 s of CPU time to analyse one matrix with 82 215 unique elements) residues *B*1–*B*11, *B*14–*B*101 and *B*108–*B*122, *i.e.* most residues of the variable domain of the heavy chain, as the largest conformationally invariant part.

Superposition of the two NCS-related conformers of the heavy chain based on the 114 conformationally invariant C^{α} atoms resulted in a mean r.m.s. deviation of 0.213 Å, which is of the order of the experimental error.

Fig. 3(*b*) clearly shows the difference in relative orientation of the constant and the variable domain arising from localized structural changes in the so-called elbow region. This movement of the domains relative to one another has been observed previously in many cases (*e.g.* Strong *et al.*, 1991; Kleywegt, 1996). In addition, there is a small but significant movement of residues *B*12 and *B*13 that is mediated by a contact of the amide group of Gln13*B* to the moving part of the elbow region. The other part of the variable half of the heavy chain that is found to be flexible is the the CDR 3 region of this chain. This region is known to be the most flexible of the loops involved in antigen recognition (Brändén & Tooze, 1999).

To fully rationalize the flexibility of this Fab fragment, the algorithm used to identify the largest conformationally invariant part should be applied iteratively to identify all rigid domains. This will be the subject of future work.

3.3. Mitochondrial aspartate aminotransferase

Mitochondrial aspartate aminotransferase from chicken is a classic example of an enzyme where binding of a substrate induces a transition from an open to a closed conformation. Following a division suggested by McPhalen, Vincent, Picot *et al.* (1992), the molecule is composed of a large (residues 48–325) and a small (residues 15–47 and 326–410) domain (Fig. 5*a*).

In an initial comparison (McPhalen, Vincent & Jansonius, 1992) of an open unliganded (PDB code 7aat) and a closed liganded form (PDB code 1ama) using the *sieve fitting procedure* (Lesk & Chothia, 1984), the conformationally invariant part of the molecule was determined to contain residues 4–12, 50–161, 167–195, 199–224 and 233–309. 2 y later, Hohenester & Jansonius (1994) presented another detailed conformational analysis of the enzyme with and without substrate, but now based on the comparison of five crystal structures in four different crystal forms. This analysis led to the conclusion that the enzyme exists in only two unique conformations (an open one and a closed one) and the domain structure found earlier was confirmed. An overview of the five structures discussed by Hohenester & Jansonius (1994) is given in Table 2.

The parameters relevant for the calculation of DPI values for the five structures are given in Table 1. Owing to the low resolution for 1tas and 1tat, the number of parameters refined is larger than the number of observables, precluding the use of (5). As there are also no $R_{\rm free}$ values available for these two refinements, (4) also cannot be used. To be able to include these two models in the analysis despite the fact that no reliable indicators of the quality of the models were available, coordinate uncertainties were calculated based on $R_{\rm free}$ values deliberately assumed to be 30.0%. The resulting values for DPI_f are rather high (0.53 and 0.54 Å for 1tas and 1tat, respectively) reflecting the rather limited information content of the two models.

The first step of the analysis concerned the identification of models representing identical conformations. Towards this objective, EDD matrices for C^{α} atoms were calculated for all possible pairs of models and the percentage of matrix elements larger than a given tolerance was evaluated (Table 3). Employing the criterion that two models are identical at a given tolerance level ε_l if for more than 98% of the elements of the respective EDD matrix the absolute value is smaller than ε_l (see §2.3), the five models could be divided into two groups using a tolerance of $\varepsilon_l = 2.0\sigma$ (Table 3). One group comprises the two models representing the open form, 7aat and 1tar and

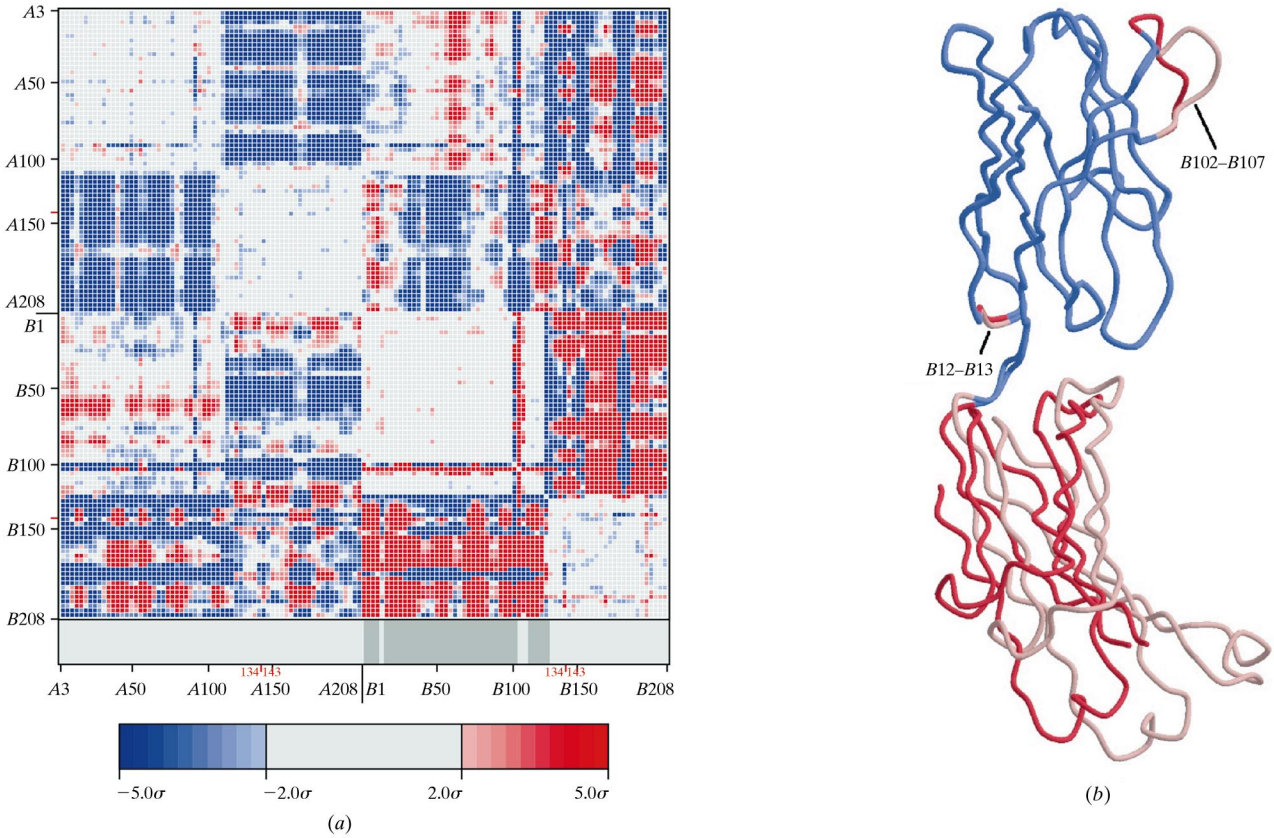

Figure 3
(*a*) Error-scaled difference distance matrix for the two NCS-related models of the Fab fragment. Residue numbers *A*3–*A*208 and *B*1–*B*208 correspond to the light and the heavy chain, respectively. Colour coding and rigid-body marking are as in Fig. 1. Red tickmarks and residue numbers correspond to stretches of residues not present in one or both of the models. The matrix underwent 3 × 3 binning before being displayed. (*b*) Superposition of the two conformers of the two heavy chains, chain *B* (blue and dark red) and chain *D* (blue and light red) of PDB entry 8fab, using the 114 conformationally invariant atoms.

the other group contains the three models representing the closed form, 1ama, 1tas and 1tat. It should be noted that the absence of significant differences between the different models for the closed form partly reflects the relatively high uncertainty assigned to the coordinates of 1tas and 1tat.

Taking the most precisely determined representative of each of the two sets, it remained to compare 7aat and 1ama. The error-scaled difference distance matrix (Fig. 4) clearly shows the division of the protein into the two domains as defined earlier. The automatic analysis of the 80 200 matrix elements using a tolerance of $\varepsilon_I = 2.0\sigma$ took 6.0 s of CPU time and flagged 281 atoms (residues 4–12, 47–226 and 232–329) as conformationally invariant. This set of conformationally invariant atoms is in very good agreement with the set defined by McPhalen and coworkers (4–12, 50–161, 167–195, 199–224 and 233–309). Apart from small variations at the boundaries of the flagged regions, only two regions, 162–166 and 196–199, located at the interface between the large and the small domain are categorized differently in the present analysis. For residues 162–166, the B values of all C^{α} atoms are above the average B value for C^{α} atoms. As a consequence, the structural differences in this region, although being substantial on an absolute scale, are not considered to be significant in the present analysis. If a second analysis is run employing a lower tolerance level of $\varepsilon_I = 1.0\sigma$, residues 196–199 (and also residues 162–166) are in fact excluded from the conformationally invariant part, indicating that the conformational changes observed are on the border of being significant.

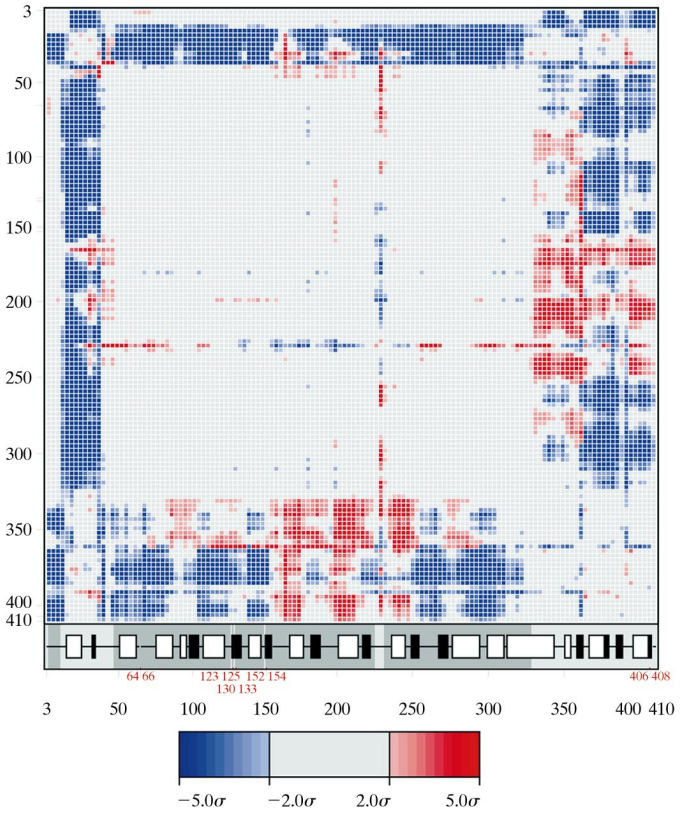

Figure 4
Error-scaled difference distance matrix for the two most precise models of aspartate aminotransferase in the open and in the closed conformation (7aat and 1ama, respectively). Colour coding, secondary-structure representation and rigid-body marking are as in Fig. 1. Red tickmarks and residue numbers correspond to stretches of residues not present in one or both of the models. The matrix underwent 4 × 4 binning before being displayed.

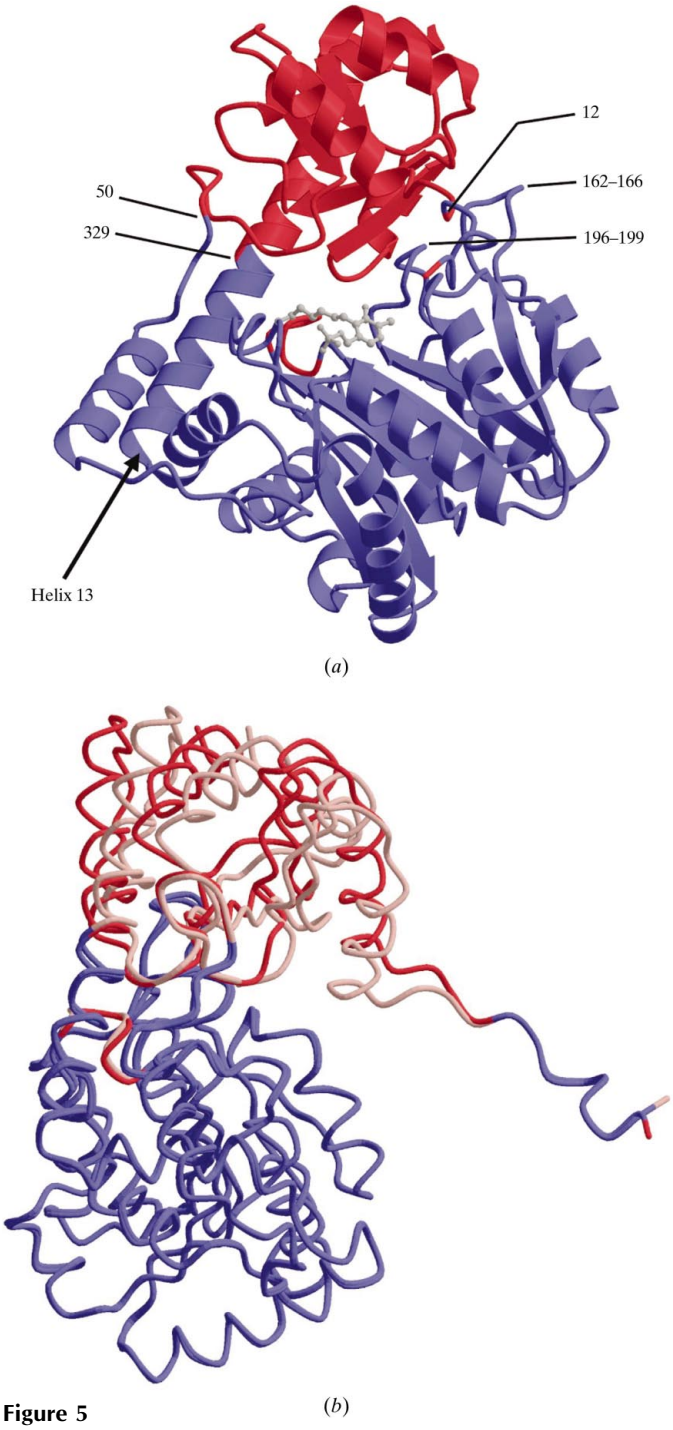

Figure 5
(*a*) Secondary structure of aspartate aminotransferase. Parts identified as conformationally invariant are shown in blue, flexible regions in red, the cofactor PLP and the lysine side chain it is bound to are shown in grey to indicate the the active site. (*b*) Superposition of 7aat (blue and dark red) and 1ama (blue and light red) using the 281 conformationally invariant atoms. This view is related to the view in (*a*) by a rotation of 90° about the vertical axis.

Table 2
Crystallographic data for AATase structures discussed in Hohenester & Jansonius (1994).

7aat and 1tar corrspond to the two open structures and 1ama, 1tas and 1tat to the three closed structures (OP1, OP2 and CL1, CL2, CL3, respectively, in Hohenester & Jansonius, 1994). 7aat supersedes the structure determined previously under the same conditions but at lower resolution by Ford *et al.* (1980). PLP stands for pyridoxal-5′-phosphate.

PDB code	Space group	Unit-cell parameters		Ligand and reference
		a, *b*, *c* (Å)	α, β, γ (°)	
7aat	*P*1	55.6, 58.7, 75.9	85.2, 109.2, 115.6	None (McPhalen, Vincent, Picot *et al.*, 1992)
1tar	*P*1	57.4, 59.4, 65.50	83.1, 104.8, 83.3	None (Hohenester & Jansonius, 1994)
1ama	$C222_1$	69.7, 91.4, 128.5	90.0, 90.0, 90.0	α-Methylaspartate (McPhalen, Vincent, Picot *et al.*, 1992)
1tas	$P2_1$	57.4, 52.4, 136.9	90.0, 101.5, 90.0	α-Methylaspartate (Hohenester & Jansonius, 1994)
1tat	$P2_12_12_1$	69.5, 89.6, 144.7	90.0, 90.0, 90.0	Maleate (Hohenester & Jansonius, 1994)

Table 3
Percentage of error-scaled difference distance matrix elements smaller than 2σ for all pairwise comparisons of 7aat, 1tar, 1ama, 1tas and 1tat.

Cases for which the percentage is larger than 98.0 are in bold.

	7aat	1tar	1ama	1tas	1tat
7aat		**100.0**	73.3	76.3	82.0
1tar			81.9	84.5	86.8
1ama				**100.0**	**100.0**
1tas					**100.0**
1tat					

large domain (residues 1–403) consisting of an eightfold α/β-barrel and a small domain (residues 404–496) folding into a compact Greek key β-barrel (Fig. 7).

The coordinate errors were estimated using (4) and (5) depending on whether or not free *R* values were available. The coordinate precision was found to be between 0.12 and 0.20 Å for the majority of the structures (Table 1). Exceptions are 1ppi and 1bvn, for which higher mean coordinate uncertainties of 0.26 and 0.37 Å are observed. In the case of 1ppi, the relatively low solvent content (46%, Table 4) results in a worse ratio of observables to parameters than for structures refined to a comparable resolution but with a higher solvent content, for example, 1pig, 1pif and 1ose with a solvent content of 72% (Table 4). For 1bvn, the low completeness of the X-ray data (probably owing to experimental difficulties connected with the relatively long *c* axis of the crystallographic unit cell) has a deleterious effect on the coordinate precision.

The largest discrepancy in categorization is found for residues 309–329, corresponding to the N-terminal half of the long α-helix (helix 13, Fig. 5) connecting the two domains. McPhalen and coworkers observed that '*the first part of helix 13 rotates by 4° and shifts by 0.4 Å*' (McPhalen, Vincent, Picot *et al.*, 1992) upon domain closure and therefore did not include this region in the rigid part of the molecule. Furthermore, they found that in the C-terminal part the helix moves around a kink angle whose centre is located in the region of residue 328–330. The movement of the first part of the helix is not detected as being significant in the present analysis. The kink region is clearly indicated by the last residue of the conformationally invariant region being identified as residue 329. In fact, in a recent analysis of the different X-ray conformers of aspartate aminotransferase using the *DYNDOM* approach, the N-terminal half of helix 13 is also included in the large rigid domain and the domain boundary is located at residue 328 (Hayward, 1999). The *DYNDOM* analysis also includes residues 162–166 and 196–199 in the rigid part of the molecule.

3.4. Pig pancreatic α-amylase

Pig pancreatic α-amylase is a sugar-cleaving enzyme consisting of 496 residues with a total molecular weight 55.3 kDa. As of January 2001, the PDB contained seven entries for porcine pancreatic amylase (E.C. 3.2.1.1, SwissProt P00690) in four different crystal forms. There is one structure of the native protein alone (1pif) and six structures with different ligands (Table 4). The molecule is composed of a

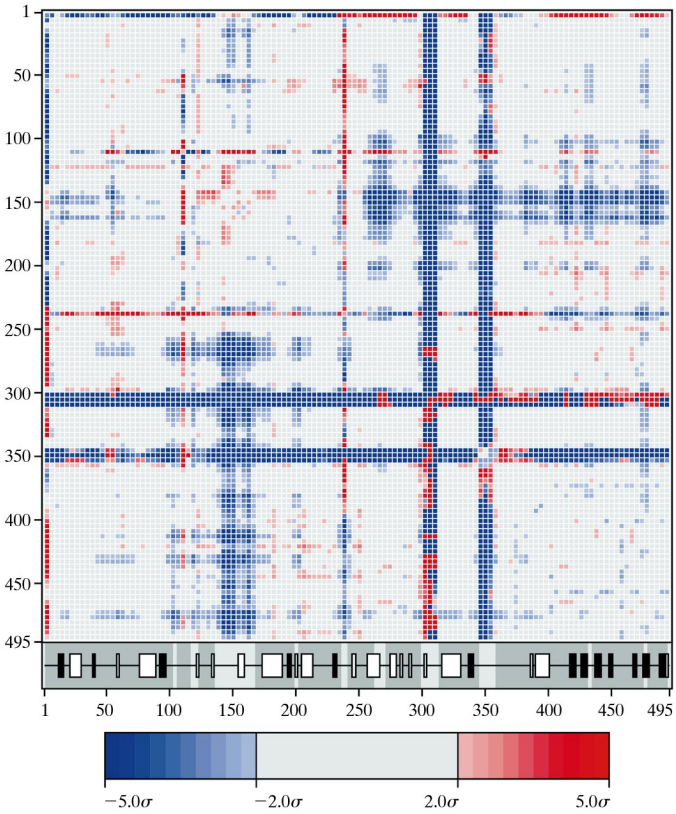

Figure 6
Matrix showing the highest absolute values found in any of the six matrices corresponding to all pairwise comparisons between the four models 1bvn, 1dhk, 1jfh and 1pig. Colour coding, secondary structure representation and rigid-body marking are as in Fig. 1.

Table 4
Crystallographic data for structures in the PDB containing pig pancreatic α-amylase (as of 31 January 2001).

All data were taken from the respective PDB files. SC, solvent content.

PDB code	Space group	Unit-cell parameters		SC (%)	d_{min} (Å)	Ligand and reference
		a, *b*, *c* (Å)	α, β, γ (°)			
1pig	$P2_12_12_1$	70.5, 114.8, 118.7	90.0, 90.0, 90.0	72	2.2	Oligosaccharide V-1532 (Machius *et al.*, 1996)
1pif	$P2_12_12_1$	70.7, 114.9, 118.9	90.0, 90.0, 90.0	72	2.3	None (Machius *et al.*, 1996)
1ose	$P2_12_12_1$	70.6, 114.7, 118.5	90.0, 90.0, 90.0	72	2.3	Acarbose (Gilles *et al.*, 1996)
1jfh	$P2_12_12_1$	56.3, 87.8, 103.4	90.0, 90.0, 90.0	45	2.03	Substrate analogue (Qian *et al.*, 1997)
1ppi	$P2_12_12_1$	56.3, 87.8, 103.4	90.0, 90.0, 90.0	46	2.2	Acarbose (Qian *et al.*, 1995)
1dhk	*C*2	151.6, 79.4, 68.0	90.0, 91.5, 90.0	56	1.85	Bean lectin-like inhibitor (Bompard-Gilles *et al.*, 1996)
1bvn	$P6_522$	77.7, 77.7, 359.5	90.0, 90.0, 120.0	51	2.5	Tendamistat (Wiegand *et al.*, 1995)

A first round of EDD-matrix analysis including all seven models revealed that four structures (1ppi, 1ose, 1pif and 1pig) fulfil the criterion for being identical (§2.3) for all their pairwise comparisons at the 2.0σ level (Table 5). Interestingly, this group contains both structures with the largest DPI values: for these two structures the absolute differences in conformation would need to be rather large to be significant. Of this group of identical structures, only one representative, namely the structure with the lowest coordinate uncertainty, 1pig, was retained for further analysis.

1pig and the remaining three models, 1bvn, 1dhk and 1jfh were then subjected to rigid-body analysis with standard parameters ($n_{hyp} = 20$, $w_p = 20.0$, $\varepsilon_l = 2.0$, $\varepsilon_h = 5.0$, $r_{mut} = 5.0\%$). The evolutionary search against six EDD matrices containing a total of 855 855 elements converged to homogeneity after five generations (corresponding to 9.4 s CPU time) marking 413 out of 495 residues as conformationally invariant. Fig. 6 shows an error-scaled difference distance matrix summarizing the six difference distance matrices calculated for the four structures. The atom set suggested as conformationally invariant by the genetic algorithm is clearly consistent with this matrix. Furthermore, a projection of the residues marked as rigid and flexible onto a schematic view of the protein produces a very convincing picture (Fig. 7). All the flexible stretches of amino acids, with the exception of residues 476–477 and the very N-terminal residue 1, are located on the substrate-binding side of the protein, indicating that the major part of the molecule forms a rigid scaffold to which loops involved in ligand binding and catalysis are attached. The small non-rigid region in the Greek key domain (476–477) reflects a crystallization artefact: in the 1pig model, a sugar molecule wedged in between two symmetry-related protein molecules causes a distortion of the polypeptide backbone in this region.

3.5. Comparions of ten NCS-related copies of epimerase

ADP-L-glycero-D-mannoheptose 6-epimerase (EC 5.1.3.20; MW = 34.9 kDa; 310 amino-acid residues) is an enzyme required for the biosynthesis of lipopolysaccharides in many pathogenic bacteria. The structure of the enzyme in complex with NADP and ADP-glucose has been determined to a resolution of 2.0 Å using monoclinic crystals (space group $P2_1$; unit-cell parameters $a = 99.5$, $b = 109.8$, $c = 181.5$ Å, $\beta = 91.0°$) by Deacon *et al.* (2000). Architecturally, the protein molecule can be subdivided into an N-terminal and a C-terminal domain (residues 1–167, 214–236 and 280–292, and residues 168–213, 237–279 and 293–310, respectively; Fig. 8*a*). The N-terminal domain binds NADP and consists of a modified Rossmann fold with a central seven-stranded β-sheet flanked on either side by a total of seven helices. The C-terminal domain consists of an arrangement of two small β-sheets and three α-helices. The asymmetric unit contains two pentamers and superposition of the ten crystallographically independent molecules revealed that the individual molecules adopt slightly different conformations depending on the crystal packing. As parts of the sequence could not be built for all of the model, the following analysis is limited to the 271 C^α atoms present in all ten molecules of the model.

An initial inspection of all pairwise comparisons to establish which conformers are identical according to the criterion defined in §2.3 did not reveal any consistent sets of identical molecules.

Running the GA with standard parameters ($n_{hyp} = 20$, $w_p = 20.0$, $\varepsilon_l = 2.0$, $\varepsilon_h = 5.0$, $r_{mut} = 5.0\%$), convergence to homogeneity was reached after six generations (20.9 s CPU time) and 205 of 271 C^α atoms were marked as conformationally invariant. The residues selected as rigid (1–21, 24–79, 86–120, 125–177, 210–238 and 278–288) fit well with the domain structure of the protein described above. One half of the protein provides a rigid scaffold in which the cofactor is firmly anchored in a way such that the nicotinamide moiety points toward the active site. The other half of the molecule consists of flexible parts that are involved in substrate binding (Fig. 8). The rigid domain consists mostly of residues from the N-terminal half of the sequence. For residues 1–177, flexible parts are only found for residues 22 and 23 (a glycine and a lysine in a surface loop region) and 80–85 and 121–124 (both loops protruding from the C-terminal towards the substrate binding site). Most of the C-terminal half of the polypeptide chain is flexible. However, the regions of the C-terminal half for which the chain returns into the N-terminal domain are rigid again: these rigid parts include residues 210–238 (adding another helix and the seventh β-sheet to the modified Rossmann fold) and residues 278–288 comprising a short helix interacting with the central β-sheet).

Employing the 205 conformationally invariant C^α atoms for least-squares superposition of all molecules onto molecule *A* gave mean r.m.s. deviations between 0.11 and 0.18 Å with a

Table 5
Percentage of EDD elements smaller than 2σ for all pairwise comparisons of seven pig pancreatic α-amylase structures. Cases for which the percentage is larger than 98.0 are shown in bold.

	1bvn	1dhk	1jfh	1ppi	1ose	1pif	1pig
1bvn		95.2	96.3	97.7	97.5	**98.8**	96.5
1dhk			88.9	93.8	90.9	92.0	89.9
1jfh				**99.5**	98.5	96.4	97.8
1ppi					**99.9**	**98.6**	**99.9**
1ose						**98.9**	**100.0**
1pif							**98.1**
1pig							

mean value of 0.14 ± 0.02 Å. These deviations are comparable to the coordinate error estimated *via* the DPI method, $\mathrm{DPI}_f = 0.17$ Å (Table 1), and are significantly smaller than the mean r.m.s. deviation of 0.39 ± 0.07 Å that was obtained when all 271 C^{α} atoms present in all molecules were superimposed.

A graphical reprensentation of the superposition of the ensemble of structures highlights the rigidity of the C-terminal domain and the flexibility of the N-terminal domain (Fig. 8*b*). The superposition also shows that the ensemble of conformers can be divided into two groups, one corresponding to a more open and the other to a more closed form (see also Deacon *et al.*, 2000).

4. Conclusions and future perspectives

A genetic algorithm for the identification of the part of a protein molecule that is conformationally invariant with respect to a set of N conformers has been designed, implemented and tested on five examples using a standard set of parameters. As a result of the acceleration provided by the application-specific genetic operators (§2.2.2), the algorithm converges rapidly and the results are in good agreement with elaborate manual analysis. Genetic algorithms are particularly suitable for this problem because hypotheses can be expressed in sequential binary form and partially correct solutions begin to have a positive influence on the fitness at an early stage.

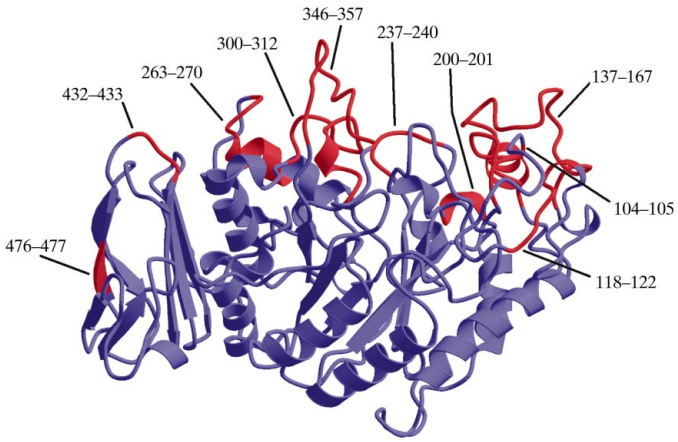

Figure 7
Secondary structure of pig pancreatic α-amylase. Parts identified as conformationally invariant (residues 2–103, 106–117, 123–136, 168–199, 202–236, 241–262, 271–299, 313–345, 358–431, 434–475 and 478–495) are shown in blue; flexible regions are shown in red.

This tool being available, the following general procedure for comparing a set of models of a molecule can be suggested: (i) identification and removal of redundant conformers, (ii) identification of a set of conformationally invariant atoms with respect to the non-redundant set of conformers, (iii) least-squares superposition of all models using the conformationally invariant subset only and (iv) graphical inspection of the results.

The first step can be performed by calculating error-scaled difference distance matrices for all pairs of conformers in the initial set. If groups of conformers turn out to be not significantly different, as indicated by a small number of non-zero elements in the corresponding error-scaled difference distance matrix (see §2.3), only the most precise conformer of each set of identical models is retained for the subsequent analysis. This will speed up the analysis and more importantly will avoid bias in the selection of the conformationally invariant atoms caused by multiple inclusion of identical models. In the future, this clustering analysis could be automated employing clustering algorithms such as that described by Kelley *et al.* (1996).

The central step of the procedure, the identification of the rigid part of the molecule by simultaneous analysis of a potentially large number of conformers, is facilitated by the genetic algorithm presented. The algorithm can analyse a large number of complex difference distance matrices very rapidly using only moderate computing resources. Moreover, such an automated procedure is more objective than an iterative manual procedure.

A standard set of parameters has been shown to work under rather different circumstances. Nevertheless, in the case that no satisfactory result is obtained with the standard parameter set, the parameters can be changed by the user of the program. One problematic scenario is the comparison of models suffering from serious systematic errors. For such cases, difference distance matrices will be very noisy even after the application of error-scaling and it may be necessary to choose rather high tolerance levels ε_l to obtain any reasonable results. On the other hand, if well determined crystal structures are used, reduction of ε_l may be beneficial in order to impose an extremely strict criterion for conformational invariance. Such lower tolerance levels will of course yield smaller sets of atoms to be used for superposition. Theoretically at least, this is not a problem, as in principle three non-collinear atoms are sufficient for a least-squares superposition of molecules in three dimensions. However, in reality a larger number of atoms is advantageous in order to perform a statistically robust superposition. The parameters governing the evolutionary search (number of hypotheses, creation of the starting population, mutation rate *etc.*) normally do not have to be changed.

The least-squares superposition itself can be performed using classical least-squares methods, such as the one implemented in *LSQKAB* (Kabsch, 1976). Such standard least-squares methods, however, do not weight the residual coordinate difference with respect to the coordinate precision

of the atoms involved, allowing the superposition to be too strongly influenced by large coordinate differences between atoms of high coordinate uncertainty. To alleviate this problem, we are working on the implementation of an uncertainty-weighted least-squares superposition algorithm. Meanwhile, one possibility is to rerun the present algorithm with a very strict tolerance criterion: this will automatically identify a subset of the set of the conformationally invariant atoms that is rigid and has a low coordinate error.

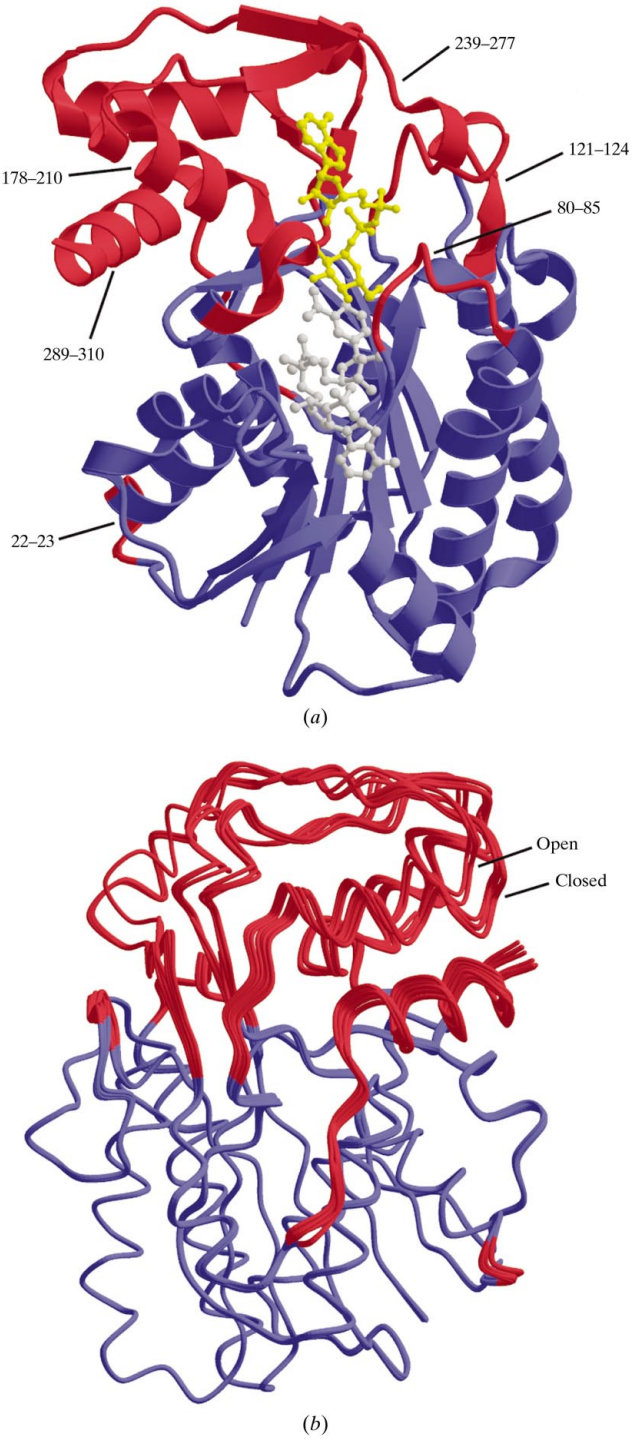

Figure 8
(*a*) Secondary structure of epimerase. Parts identified as conformationally invariant are shown in blue and flexible regions in red. The cofactor NADP and the substrate analogue ATP-glucose are shown in ball-and-stick representation in grey and yellow, respectively. (*b*) Backbone traces of ten molecules superimposed using the conformationally invariant part (shown in blue). This view is related to the view in (*a*) by a rotation of about 120° about the vertical axis.

Finally, displaying the results on a computer graphics system in a suitable format will in many cases allow intuitive interpretation of the results.

The entire procedure crucially depends on accurate error-estimates and, of course, on the models employed being essentially correct and the associated statistics being reliable. Although Cruickshank's DPI gives a good first-order approximation, it would be desirable to have more accurate estimators of coordinate precision. A more complex functional form with empirically optimized parameters could be an improvement (Cruickshank, 1999). Ideally, coordinate uncertainties would be calculated directly from the experimental data, an obvious choice being an estimate based on real-space correlation coefficients (Jones *et al.*, 1991). Technically, such data-based estimates require the diffraction data to be available at the time of the calculation, which unfortunately is not always the case, as diffraction data are not always deposited together with the model coordinates (Jiang *et al.*, 1999).

So far, the method has only been applied to C^{α} atoms as 'representatives' of overall protein conformation. The algorithm is equally well applicable to, for example, P atoms in RNA structures or to the study of the changes in relative positions of atoms in an active site when different substrates or inhibitors are bound.

Furthermore, iterative application of the method would allow the delineation of not only the largest but also smaller domains in order to characterize multi-domain structures both in terms of the identification of the domains and the description of their relative motions (Verbitsky *et al.*, 1999; Gerstein & Krebs, 1998; Berendsen & Hayward, 1998). This will be the subject of future work.

Currently, the method is limited to the analysis of different conformers of *the same* molecule. A generalization of the algorithm for cases where closely homologeous structures are compared and a robust mapping of corresponding atoms in three dimensions can be achieved using existing methods such as *DALI* (Holm & Sander, 1993), which is also based on difference distance matrices, is a mostly technical problem [for recent reviews concerning structure alignment see, for example, Lemmen & Lengauer (2000) and Eidhammer *et al.* (2000)]. Such a superposition of homologeous models taking the different level of precision into account would not only be useful from a structure-analysis point of view but could also be used to construct hybrid models for use in difficult molecular-replacement cases as suggested by Read (2001).

I am grateful to George Sheldrick for many discussions and encouragement, and to Karl Edman for being a patient and creative tester of early versions of the computer program implementing the ideas presented. The computer program *ESCET* is available as a beta-test version from the author.

References

Berendsen, H. J. & Hayward, S. (1998). *Proteins Struct. Funct. Genet.* **30**, 144–154.
Bompard-Gilles, C., Rousseau, P., Rouge, P. & Payan, F. (1996). *Structure*, **4**, 1441–1452.
Brändén, C. & Tooze, J. (1999). *Introduction to Protein Structure.* London: Garland Publishing.
Chacón, P., Díaz, J. F., Morán, F. & Andreu, J. M. (2000). *J. Mol. Biol.* **299**, 1289–1302.
Chang, G. & Lewis, M. (1994). *Acta Cryst.* D**50**, 667–674.
Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.
Deacon, A. M., Ni, Y. S., Coleman, W. G. & Ealick, S. (2000). *Structure*, **15**, 453–462.
Eidhammer, I., Jonassen, I. & Taylor, W. R. (2000). *J. Comput. Biol.* **7**, 685–716.
Ford, G. C., Eichele, G. & Jansonius, J. N. (1980). *Proc. Natl Acad. Sci. USA*, **77**, 2559–2563.
Gerstein, M. & Altman, R. B. (1995). *J. Mol. Biol.* **252**, 151–175.
Gerstein, M. & Krebs, W. (1998). *Nucleic Acids Res.* **26**, 4280–4290.
Gilles, C., Astier, J. P., Marchis-Mouren, G., Cambillau, C. & Payan, F. (1996). *Eur. J. Biochem.* **238**, 562–559.
Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Boston, MA, USA: Addison Wesley.
Hayward, S. (1999). *Proteins Struct. Funct. Genet.* **36**, 425–435.
Hohenester, E. & Jansonius, J. N. (1994). *J. Mol. Biol.* **236**, 963–968.
Holland, J. H. (1975). *Adaption in Natural and Artificial Systems.* University of Michigan Press.
Holm, L. & Sander, C. (1993). *J. Mol. Biol.* **233**, 123–138.
Jiang, J., Abola, E. & Sussman, J. L. (1999). *Acta Cryst.* D**55**, 4.
Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). *J. Mol. Biol.* **267**, 727–748.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
Kelley, L., Gardner, S. & Sutcliffe, M. J. (1996). *Protein Eng.* **9**, 1063–1065.
Kelley, L., Gardner, S. & Sutcliffe, M. J. (1997). *Protein Eng.* **10**, 737–741.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Kleywegt, G. J. (1996). *Acta Cryst.* D**52**, 842–857.
Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.
Ladner, J. E., Reddy, P., Davis, A., Tordova, M., Howard, A. & Gilliland, G. L. (2000). *Acta Cryst.* D**56**, 673–683.
Lemmen, C. & Lengauer, T. (2000). *J. Comput. Aided Mol. Des.* **14**, 215–232.
Lesk, A. M. & Chothia, C. (1984). *J. Mol. Biol.* **174**, 175–191.
McPhalen, C. A., Vincent, M. G. & Jansonius, J. N. (1992). *J. Mol. Biol.* **225**, 495–517.
McPhalen, C. A., Vincent, M. G., Picot, D., Jansonius, J. N., Lesk, A. M. & Chothia, C. (1992). *J. Mol. Biol.* **227**, 197–213.
Machius, M., Vertesy, L., Huber, R. & Wiegand, G. (1996). *J. Mol. Biol.* **260**, 409–421.
Merritt, E. & Murphy, M. (1994). *Acta Cryst.* D**50**, 869–873.
Mitchell, M. (1996). *An Introduction to Genetic Algorithms.* Cambridge, MA, USA: The MIT Press.
Nichols, W. L., Rose, G. D., Ten Eyck, L. & Zimm, B. H. (1995). *Proteins Struct. Funct. Genet.* **23**, 38–48.
Nichols, W. L., Zimm, B. H. & Ten Eyck, L. (1997). *J. Mol. Biol.* **270**, 598–615.
Qian, M., Haser, R. & Payan, F. (1995). *Protein Sci.* **4**, 747–755.
Qian, M., Spinelli, S., Driguez, H. & Payan, F. (1997). *Protein Sci.* **6**, 2285–2296.
Read, R. (2001). *Acta Cryst.* D**57**, 1373–1382.
Saul, F. A. & Poljak, R. J. (1992). *Proteins*, **14**, 363–371.
Schmidt, R., Gerstein, M. & Altman, R. B. (1997). *Protein Sci.* **6**, 246–248.
Schneider, T. R. (2000). *Acta Cryst.* D**56**, 714–721.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Strong, R. K., Petsko, G. A., Sharon, J. & Margolies, M. N. (1991). *Biochemistry*, **30**, 3749–3757.
Verbitsky, G., Nussinov, R. & Wolfson, H. (1999). *Proteins*, **34**, 232–254.
Webster, G. & Hilgenfeld, R. (2001). *Acta Cryst.* A**57**, 351–358.
Wiegand, G., Epp, O. & Huber, R. (1995). *J. Mol. Biol.* **247**, 99–110.
Wriggers, W. & Schulten, K. (1997). *Proteins Struct. Funct. Genet.* **29**, 1–14.