HermiteFit: fast-fitting atomic structures into a low-resolution density map using three-dimensional orthogonal Hermite functions

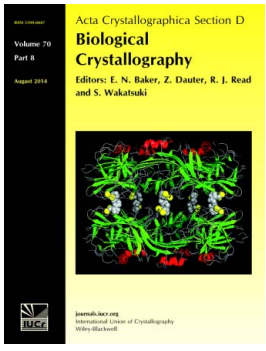
Georgy Derevyanko, Sergei Grudinin

Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

HermiteFit: fast-fitting atomic structures into a low-resolution density map using three-dimensional orthogonal Hermite functions

Georgy Derevyanko and Sergei Grudinin

Acta Cryst. (2014). D**70**, 2069–2084



Acta Crystallographica Section D: Biological Crystallography welcomes the submission of papers covering any aspect of structural biology, with a particular emphasis on the structures of biological macromolecules and the methods used to determine them. Reports on new protein structures are particularly encouraged, as are structure–function papers that could include crystallographic binding studies, or structural analysis of mutants or other modified forms of a known protein structure. The key criterion is that such papers should present new insights into biology, chemistry or structure. Papers on crystallographic methods should be oriented towards biological crystallography, and may include new approaches to any aspect of structure determination or analysis. Papers on the crystallization of biological molecules will be accepted providing that these focus on new methods or other features that are of general importance or applicability.

Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

HermiteFit: fast-fitting atomic structures into a low-resolution density map using three-dimensional orthogonal Hermite functions

Georgy Derevyanko[a,b,c,d,e] and Sergei Grudinin[f,g,h]*

[a]Institute of Complex Systems (ICS-6), Forschungszentrum Jülich, Jülich, Germany, [b]Université Grenoble Alpes, IBS, F-38000 Grenoble, France, [c]CNRS, IBS, F-38000 Grenoble, France, [d]CEA, IBS, F-38000 Grenoble, France, [e]Research-Educational Centre Bionanophysics, Moscow Institute of Physics and Technology, Dolgoprudniy, Russia, [f]Université Grenoble Alpes, LJK, F-38000 Grenoble, France, [g]CNRS, LJK, F-38000 Grenoble, France, and [h]INRIA, France

Correspondence e-mail: sergei.grudinin@inria.fr

Received 21 January 2013
Accepted 19 May 2014

HermiteFit, a novel algorithm for fitting a protein structure into a low-resolution electron-density map, is presented. The algorithm accelerates the rotation of the Fourier image of the electron density by using three-dimensional orthogonal Hermite functions. As part of the new method, an algorithm for the rotation of the density in the Hermite basis and an algorithm for the conversion of the expansion coefficients into the Fourier basis are presented. *HermiteFit* was implemented using the cross-correlation or the Laplacian-filtered cross-correlation as the fitting criterion. It is demonstrated that in the Hermite basis the Laplacian filter has a particularly simple form. To assess the quality of density encoding in the Hermite basis, an analytical way of computing the crystallographic R factor is presented. Finally, the algorithm is validated using two examples and its efficiency is compared with two widely used fitting methods, *ADP_EM* and *colores* from the *Situs* package. *HermiteFit* will be made available at http://nano-d.inrialpes.fr/software/HermiteFit or upon request from the authors.

1. Introduction

An important class of algorithms in computer science deals with the exhaustive search in six-dimensional space of translations and rotations of a rigid body. These algorithms are used, for example, in crystallography for molecular replacement and in computational biology to perform ligand docking, to predict protein–protein interactions and to discover the structures of macromolecular assemblies.

Modern exhaustive search algorithms implement either a fast three-dimensional translational search using the fast Fourier transform (FFT; Chacón & Wriggers, 2002; Katchalski-Katzir *et al.*, 1992; Gabb *et al.*, 1997; Wriggers, 2010; Siebert & Navaza, 2009) or a fast three-dimensional rotational search by means of spherical harmonics decomposition and the FFT (Kovacs & Wriggers, 2002; Crowther, 1972), or even a fast five-dimensional rotational search (Kovacs *et al.*, 2003; Ritchie *et al.*, 2008). An exhaustive search is also widely used as a preliminary step preceding local search or flexible refinement procedures. Thus, the quality and the speed of exhaustive search algorithms have a great impact on the solution of a vast variety of problems. Therefore, we believe that new directions of research on this topic are very important and highly valuable.

In this paper, we present the new *HermiteFit* algorithm that uses orthogonal Hermite functions to perform an exhaustive search in the six-dimensional space of rigid-body motions. We apply this method to the problem of fitting of a

high-resolution X-ray structure of a protein subunit into the cryo-electron microscopy (cryo-EM) density map of a protein complex. As part of this new method, we developed an algorithm for the rotation of the decomposition in the Hermite basis and another algorithm for the conversion of the Hermite expansion coefficients into the Fourier basis.

The choice of the application of our algorithm is dictated by the fact that currently the major source of information on the mechanisms of function of proteins and their assemblies are the atomic structures obtained by X-ray crystallography. However, as the size of the protein grows, as often happens in protein complexes, it becomes more difficult to obtain well ordered crystals that are sufficiently large for X-ray experiments. Hopefully, in many cases, different parts of the protein complex can be crystallized separately. Usually, their structures can be solved to atomic resolution. The whole protein complex in this case can be probed by cryo-EM (Cheng & Walz, 2009), by small-angle X-ray scattering (Svergun & Koch, 2003) and with recent advances in femtosecond X-ray lasers (Chapman *et al.*, 2011). Usually, these techniques provide of an electron-density map (EDM) of a large protein or a protein complex with a resolution lower than 3.5 Å, whereas the atomic structures of its small subunits can be solved with X-ray crystallography at even sub-angstrom resolution. To reconstruct large proteins or protein complexes at high resolution, the high-resolution crystallographic structures of small units can be fitted into the low-resolution structures of the whole assembly. A number of software packages have been developed for this task. The most notable of them are *Situs* (Wriggers, 2010; Chacón & Wriggers, 2002), *NORMA* (Suhre *et al.*, 2006), *EMFit* (Rossmann *et al.*, 2001) and *UROX* (Siebert & Navaza, 2009). Despite the differences in their implementation, all of the algorithms maximize some score that shows the goodness of the fitting using a certain optimization algorithm. An excellent review of the different types of scoring functions used for cryo-EM density fitting is given by Vasishtan & Topf (2011). According to them, one of the most popular scoring functions is the cross-correlation function (CCF) between the EDM and the density of the fitted protein.

Given a protein structure that is described by its electron density $f(\mathbf{r})$, and an EDM obtained from, for example, a cryo-EM experiment described by the function $g(\mathbf{r})$, we can minimize the square-root discrepancy between them. Precisely, this discrepancy is given by

$$S = \int [\hat{T}\hat{R}f(\mathbf{r}) - g(\mathbf{r})]^2 \, \mathrm{d}\mathbf{r}, \tag{1}$$

where $\hat{T}$ and $\hat{R}$ are the operators of the translation and the rotation, respectively, applied to the density $f(\mathbf{r})$. We can rewrite the scoring function S as

$$S = \int [\hat{T}\hat{R}f(\mathbf{r})]^2 \, \mathrm{d}\mathbf{r} + \int g^2(\mathbf{r}) \, \mathrm{d}\mathbf{r} - 2 \int \hat{T}\hat{R}f(\mathbf{r}) g(\mathbf{r}) \, \mathrm{d}\mathbf{r}. \tag{2}$$

Therefore, minimization of the score S is equivalent to maximization of the CCF,

$$\mathrm{CCF} = \int \hat{T}\hat{R}f(\mathbf{r}) g(\mathbf{r}) \, \mathrm{d}\mathbf{r}, \tag{3}$$

with respect to the parameters of the operators $\hat{T}$ and $\hat{R}$. This scoring function has been used in the majority of the algorithms and software packages that perform fitting to the EDM (Wriggers, 2010; Siebert & Navaza, 2009; Suhre *et al.*, 2006).

Another widely used scoring function is the Laplacian-filtered cross-correlation function (LCCF). It originated from the observation that a human performing a manual fitting of a structure into an EDM tends to match the isosurfaces of the densities rather than the densities themselves,

$$\mathrm{LCCF} = \int [\Delta \hat{T}\hat{R}f(\mathbf{r})][\Delta g(\mathbf{r})] \, \mathrm{d}\mathbf{r}. \tag{4}$$

This scoring function works better than the CCF for low-resolution maps (∼10–30 Å; Wriggers, 2010) and was used for the first time in the *CoAn*/*CoFi* algorithm (Volkmann & Hanein, 1999). Other scoring functions that, for example, penalize symmetry-induced protein–protein contacts, or make use of protein–protein docking potentials *etc.*, have also been developed (Vasishtan & Topf, 2011). In our work, we use the CCF and LCCF to determine the goodness of fit.

In this paper, we demonstrate the ability of our algorithm to compete with the well established approaches by using two examples of different difficulty: the PniB conotoxin peptide and the GroEL complex. The first example illustrates the encoding principles and demonstrates the influence of the encoding quality on the goodness of fit. The second example is the gold standard of all electron-density map-fitting algorithms. Our approach allows analytical assessment of the quality of encoding of the Hermite basis using an estimation of the crystallographic R factor. We then compare this estimation with that computed numerically for the PniB conotoxin density map. Finally, we compare the speed and the fitting accuracy of our algorithm with two popular programs, the *ADP_EM* fitting method and the *colores* program from the *Situs* package, and demonstrate that *HermiteFit* takes less running time per search point compared with the two other methods while attaining a similar accuracy.

The *HermitFit* algorithm can be straightforwardly applied to a broad class of problems in different fields of research. For example, one of the bottlenecks of algorithms for molecular replacement in crystallography is the computation of the Fourier coefficients (structure factors) of a molecule (Navaza & Vernoslova, 1995). This operation needs to be precise and fast. However, exact analytical evaluation of the structure factors is too costly (Sayre, 1951) when recomputing them for each rotation of the molecule. Therefore, currently one uses the Sayre–Ten Eyck approach to compute the Fourier coefficients (Ten Eyck, 1977). Unfortunately, one has to be very careful tuning the parameters of the electron-density model and the grid cell size to obtain the desired precision (Navaza, 2002; Afonine & Urzhumtsev, 2004). Unlike the Sayre–Ten Eyck approach, our algorithm offers an analytical expression for the structure factors of the Hermite decomposition of a molecule. Finally, our approach allows analytical estimatation of the quality of encoding using, for example, crystallographic R factors.

2. Methods

2.1. Summary of the standard fitting algorithm

The standard FFT-based three-dimensional fitting algorithm operates according to the workflow shown in Fig. 1 (Katchalski-Katzir *et al.*, 1992; Gabb *et al.*, 1997; Chacón & Wriggers, 2002). The input of this algorithm is a protein atomic structure determined experimentally by, for example, X-ray crystallography or nuclear magnetic resonance (NMR) experiments. Another input is an experimental EDM determined by means of, for example, cryo-EM. Firstly, the algorithm decomposes the experimental EDM into the Fourier basis using the fast Fourier transform algorithm. It then rotates the protein structure to a certain orientation **r** and decomposes the electron density of the rotated structure into the Fourier basis. The electron density is typically computed as a sum of Gaussians centred on non-H atoms of the protein. Afterwards, the algorithm exhaustively explores translational degrees of freedom of the rotated protein with respect to the EDM. For every translation **t**, it determines the corresponding score, which is usually given by the correlation between the two densities. This procedure is equivalent to computing the convolution of two functions,

$$\mathrm{CCF}(\mathbf{r}, \mathbf{t}) = \int f(\mathbf{r}, \mathbf{x} - \mathbf{t}) g(\mathbf{x})\, \mathrm{d}\mathbf{x}, \qquad (5)$$

where $f(\mathbf{r}, \mathbf{x} - \mathbf{t})$ is the density of the protein rotated by **r** and translated by **t** and $g(\mathbf{x})$ is the experimental electron-density map. To speed up this step, the algorithm computes the values of the Fourier transform of the CCF for all translational degrees of freedom at once using the convolution theorem. Finally, the algorithm computes the inverse Fourier transform (IFT) of the convolution, generates a new rotation of the protein structure and returns to the second step. This procedure is repeated until all rotational degrees of freedom of the protein with respect to the EDM have been explored (see Fig. 1). The solution of the fitting problem is then given by $(\mathbf{r}_{\max}, \mathbf{t}_{\max}) = \mathrm{argmax}_{\mathbf{r},\mathbf{t}}[\mathrm{CCF}(\mathbf{r}, \mathbf{t})]$.

The bottleneck of the standard algorithm is the re-projection of the protein electron density into the Fourier space after each rotation. To overcome this, we propose encoding the electron density of the protein structure in the orthogonal Hermite basis prior to performing the rotational search. This allows the projection of the protein density into the Fourier space to be sped up. Since only members of the Fourier family of linear transforms can replace the $O(N^2)$ operations of a convolution in a time domain by $O(N)$ operations in a frequency domain (Stone, 1998), we still need to perform the convolution in the Fourier space. Fig. 2 shows the workflow of the proposed algorithm. The computational complexity of this algorithm is listed in Table 1.

Table 1
Complexity of the Hermite fitting algorithm.

Here, M denotes the order of the Fourier decomposition, N is the order of the Hermite decomposition, N_{atom} is the number of atoms in the protein and N_{rot} is the number of rotations to be sampled.

Operation	Complexity	Loop multiplier
Decomposition of the step function	$O(M^3 \log M^3)$	1
Decomposition of the Gaussian	$O(N_{\mathrm{atoms}} N^3)$	1
Construction of the rotation matrix	$O(N_{\mathrm{rot}} N^4)$	1
Rotation	$O(N_4)$	N_{rot}
Evaluation of the Hermite series	$O(M^3 N + M^2 N^2 + M N^3)$	N_{rot}
Multiplication	$O(M^3)$	N_{rot}
Inverse Fourier transform	$O(M^3 \log M^3)$	N_{rot}

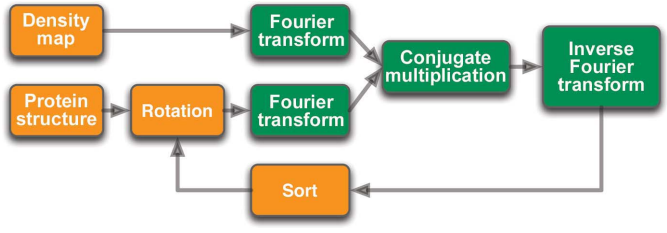


Figure 1
Flowchart of the standard fitting algorithm based on Fourier correlations. Green blocks correspond to operations in Fourier space.

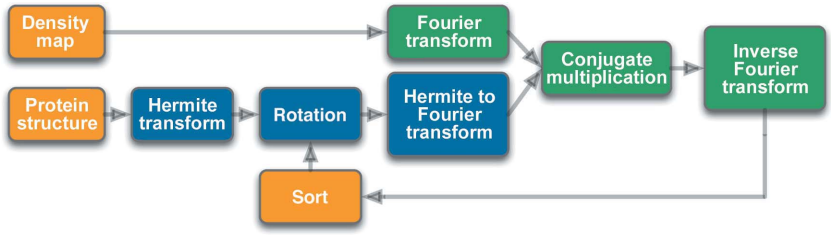


Figure 2
Flowchart of *HermiteFit*, the new fitting algorithm based on Hermite expansions. Green blocks correspond to operations in Fourier space. Blue blocks correspond to operations in Hermite space.

2.2. Hermite functions

The orthogonal Hermite function of order n is defined as

$$\psi_n(x;\lambda) = \frac{\lambda^{1/2}}{(2^n n!\pi^{1/2})^{1/2}} \exp\left(-\frac{\lambda^2 x^2}{2}\right) H_n(\lambda x), \qquad (6)$$

where $H_n(x)$ is the Hermite polynomial and λ is the scaling parameter. In Fig. 3 we show several orthogonal Hermite functions of different orders with different parameters λ. These functions form an orthonormal basis set in $L^2(\mathbb{R})$. A one-dimensional function $f(x)$ decomposed into the set of one-dimensional Hermite functions up to order N is given by

$$f(x) = \sum_{i=0}^{N} \hat{f}_i \psi_i(x;\lambda). \qquad (7)$$

Here, $\hat{f}_i$ are the decomposition coefficients, which can be determined from the orthogonality of the basis functions $\psi_i(x;\lambda)$. Decomposition of (7) is called band-limited decomposition with $\psi_i(x;\lambda)$ basis functions. To decompose the EDM and the protein structures, we employ the three-dimensional Hermite functions

$$\psi_{n,l,m}(x,y,z;\lambda) = \psi_n(x;\lambda)\psi_l(y;\lambda)\psi_m(z;\lambda), \qquad (8)$$

which form an orthonormal basis set in $L^2(\mathbb{R}^3)$. A function $f(x,y,z)$ represented as a band-limited expansion in this basis is given by

$$f(x,y,z) = \sum_{i=0}^{N} \sum_{j=0}^{N-i} \sum_{k=0}^{N-i-j} \hat{f}_{i,j,k} \psi_{i,j,k}(x,y,z;\lambda). \qquad (9)$$

2.3. Decomposition of electron densities into the orthogonal Hermite basis

One of the advantages of the orthogonal Hermite basis is that we can derive the exact analytical expression for the decomposition coefficients of a molecular structure. This allows the exact decompositions to rapidly be obtained without costly numerical integration over three-dimensional space. In our algorithm, the electron density of the protein [$f(x)$ in equation 5, upon which rotation and translation operators act] is expanded in the Hermite basis using the Gaussian model. More precisely, we model the electron density of a single atom in the molecular structure as a Gaussian centred at the atomic position $\mathbf{r}_0^{(i)}$ with the squared variance equal to $\alpha^2/2$. The electron density of the whole molecular structure is then given by the following sum,

$$M(\mathbf{r}) = \sum_{i=1}^{N_{\text{atoms}}} \exp[-|\mathbf{r} - \mathbf{r}_0^{(i)}|^2/\alpha^2], \qquad (10)$$

where $\mathbf{r}_0^{(i)}$ is the position of the ith atom, $\alpha/2^{1/2}$ is the variance of the Gaussian distribution and $\mathbf{r} = (x,y,z) \in \mathbb{R}^3$ is the sampling volume. Normally, each Gaussian should be weighted with a coefficient corresponding to the electron distribution of a particular atom. However, we omit the weights in our approximation. In Appendix *A*, we provide analytical expressions (equations 48 and 54) for the decomposition coefficients of $M(\mathbf{r})$ in the one-dimensional and the three-dimensional cases.

2.4. Laplacian filter in the Hermite basis

For medium- to low-resolution maps the Laplacian-filtered cross-correlation function gives a better match compared with the CCF (Wriggers, 2010). In the Hermite basis, the Laplacian filter has a particularly simple form. Using the well known recurrence relation for the derivatives of Hermite functions, we can easily derive the following relation for the second derivative of a one-dimensional basis function:

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi_n(x;\lambda) = \frac{\lambda^2}{2}\{[n(n-1)]^{1/2}\psi_{n-2}(x;\lambda) + (2n+1)\psi_n(x;\lambda) + [(n+1)(n+2)]^{1/2}\psi_{n+2}(x;\lambda)\}. \qquad (11)$$

A similar relationship holds for the coefficients of the decomposition:

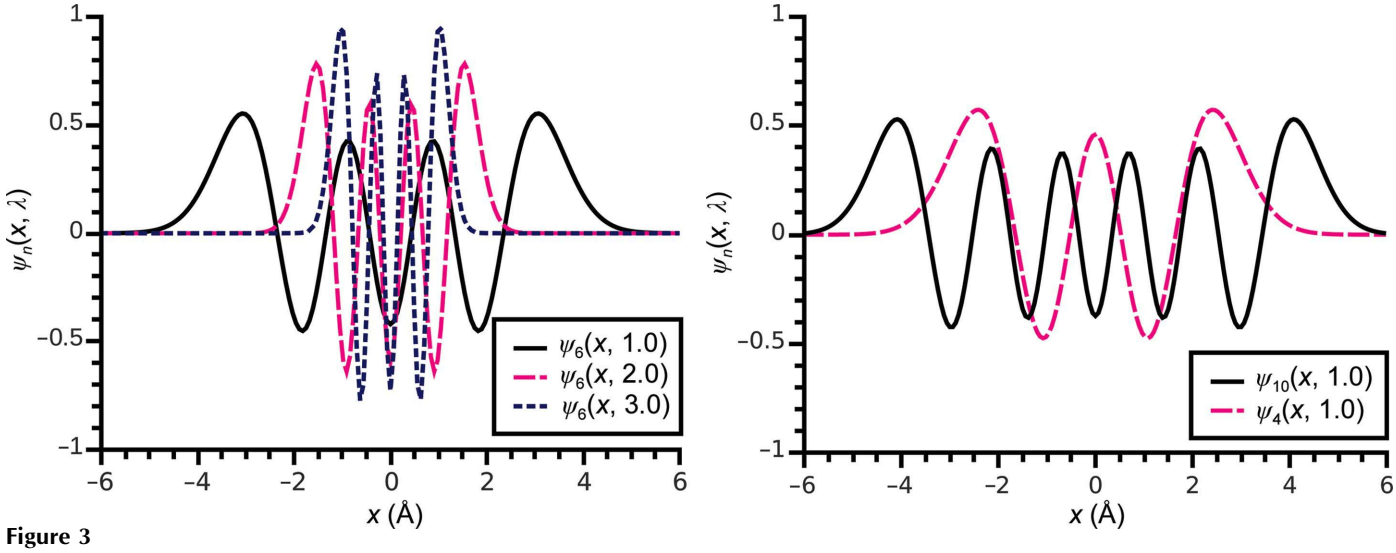


Figure 3
Left, one-dimensional Hermite functions of order six for three different scaling parameters λ. Right, one-dimensional Hermite functions of two different orders for the scaling parameter $\lambda = 1$.

$$\hat{\boldsymbol{h}}_n'' = \frac{\lambda^2}{2}\{[n(n-1)]^{1/2}\hat{\boldsymbol{h}}_{n-2} + (2n+1)\hat{\boldsymbol{h}}_n + [(n+2)(n+1)]^{1/2}\hat{\boldsymbol{h}}_{n+2}\}, \quad (12)$$

where $\hat{\boldsymbol{h}}_n$ and $\hat{\boldsymbol{h}}_n''$ are the nth-order decomposition coefficients of the original basis and its Laplacian representation, respectively. For $n < 0$ and $n > N$ we let $\hat{\boldsymbol{h}}_n = 0$ and $\hat{\boldsymbol{h}}_n'' = 0$. Owing to the properties of the Laplace operator and the three-dimensional Hermite decomposition, the contributions of the derivatives along each axis are additive. The derivation of the formula for the three-dimensional decomposition derivative is straightforward and we omit it for brevity.

2.5. Rotation of the Hermite decomposition

Recently, Park *et al.* (2009) presented a method to perform an in-plane rotation of a two-dimensional orthogonal Hermite band-limited decomposition. Here, we extend their method to the three-dimensional case. Let us first consider the decomposition of a two-dimensional function into a two-dimensional orthogonal Hermite-function basis,

$$f(x, y) = \sum_{n=0}^{N}\sum_{m=0}^{N-m}\hat{f}_{n,m}\psi_n(x;\lambda)\psi_m(y;\lambda). \quad (13)$$

The decomposition of a function $f^{\theta}(x, y)$ rotated clockwise by an angle θ is given by

$$f^{\theta}(x, y) = \sum_{m=0}^{N}\sum_{k=0}^{m}\left(\sum_{n=0}^{m}\hat{f}_{n,m-n}S_{k,n}^{m}\right)\psi_k(x;\lambda)\psi_{m-k}(y;\lambda), \quad (14)$$

where the coefficients $S_{k,n}^{m}$ are computed using the following recurrent formulae (Park *et al.*, 2009):

$$\begin{aligned}
S_{q,n}^{m+1} &= \left(\frac{n}{m-q+1}\right)^{1/2}\sin(\theta)S_{q,n-1}^{m} + \left(\frac{m-n+1}{m-q+1}\right)^{1/2}\cos(\theta)S_{q,n}^{m},\\
S_{q,0}^{m+1} &= \left(\frac{m+1}{m-q+1}\right)^{1/2}\cos(\theta)S_{q,0}^{m},\\
S_{m+1,n}^{m+1} &= \left(\frac{n}{m+1}\right)^{1/2}\cos(\theta)S_{m,n-1}^{m} - \left(\frac{m-n+1}{m+1}\right)^{1/2}\sin(\theta)S_{m,n}^{m},\\
S_{m+1,0}^{m+1} &= -\sin(\theta)S_{m,0}^{m}. \quad (15)
\end{aligned}$$

The key idea that allows the generalization of these formulae to a three-dimensional decomposition is that we can factorize a rotation in three-dimensional space into three independent in-plane rotations about three different axes and then rotate each two-dimensional decomposition using (14). Let us consider the following three-dimensional decomposition:

$$f(x, y, z) = \sum_{n=0}^{N}\psi_n(x;\lambda)\sum_{m=0}^{N-n}\sum_{l=0}^{N-m-n}\hat{f}_{n,m,l}\psi_m(y;\lambda)\psi_l(z;\lambda). \quad (16)$$

If we rotate this decomposition about the x axis, this rotation will be equivalent to N rotations of different two-dimensional decompositions in the yz plane,

$$f_n(y, z) = \sum_{m=0}^{N-n}\sum_{l=0}^{N-m-n}\hat{f}_{n,m,l}\psi_m(y;\lambda)\psi_l(z;\lambda). \quad (17)$$

This observation means that in order to perform such a rotation, we need to recompute a rank 3 tensor of coefficients $\hat{f}_{n,m,l}$ slice by slice N times using (14). Fig. 4 illustrates three subsequent rotations of the tensor $\hat{f}_{n,m,l}$. Each rotation of the coefficients in one plane corresponds to a multiplication of these coefficients by a rotation matrix. Therefore, a three-dimensional rotation defined with three Euler angles is equivalent to three sequential rotations of coefficients in three planes.

2.6. Transition from the Hermite to the Fourier basis

In order to perform a fast convolution as in (5), we convert the decomposition coefficients from the Hermite basis into the Fourier basis. This allows use of the fast convolution algorithm based on the Fourier convolution theorem, which was first introduced in protein–protein docking studies (Katchalski-Katzir *et al.*, 1992; Gabb *et al.*, 1997) and then also applied to EDM fitting (Chacón & Wriggers, 2002; Wriggers, 2010; Siebert & Navaza, 2009). The key idea of this algorithm is to compute the Fourier transform of the values of a scoring function on a grid, $\mathrm{CCF}(\mathbf{r}, \mathbf{t}) = \int f(\mathbf{r}, \mathbf{x})g(\mathbf{r}, \mathbf{x}-\mathbf{t})\,\mathrm{d}\mathbf{x}$, using the convolution theorem

$$F(f * g) = \overline{F}(f)F(g), \quad (18)$$

i.e. by multiplying the complex-conjugated coefficients of the Fourier transform of the protein electron density with the coefficients of the Fourier transform of the EDM. We then obtain $\mathrm{CCF}(\mathbf{r}, \mathbf{t})$ by taking the inverse Fourier transform of $F(f * g)$,

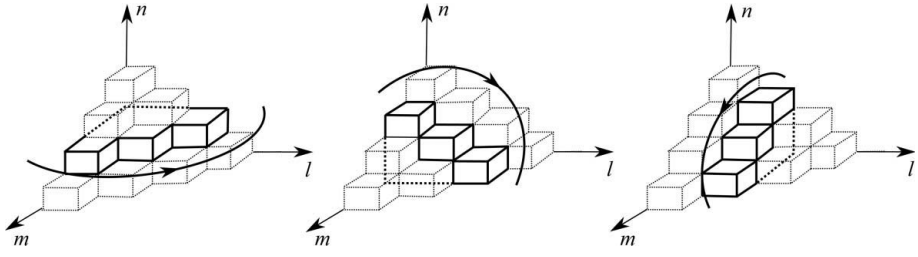


Figure 4
Sequential rotations of coefficients $\hat{f}_{n,m,l}$ about different axes. The rotated layer is shown with solid cubes; other coefficients are shown with dashed cubes. To perform the complete rotation of the decomposition about one axis, we rotate each layer of coefficients about the corresponding axis in the coefficient space.

$$\mathrm{CCF}(\mathbf{r}, \mathbf{t}) = \mathrm{IFT}[\overline{F}(f)F(g)]. \quad (19)$$

Now we explain how we convert the decomposition coefficients from the Hermite basis into the Fourier basis. Consider the decomposition of a function $f(\mathbf{r})$ in the three-dimensional Hermite basis with decomposition coefficients $\hat{f}_{i,j,k}$ (9). The orthogonal Hermite functions are the eigenfunctions of the continuous Fourier transform,

$$\int \psi_n(x;\lambda)\exp(-2\pi i\omega x)\,\mathrm{d}x = (-i)^n \psi_n\left(\omega;\frac{2\pi}{\lambda}\right) \equiv \tilde{\psi}_n(\omega;\lambda), \quad (20)$$

where ω is the frequency in reciprocal space. In order to compute Fourier coefficients of $f(\mathbf{r})$ to order M, we first compute the Fourier transforms of the basis functions $\psi_i(x;\lambda)$, $\psi_j(y;\lambda)$ and $\psi_k(z;\lambda)$ using (20). We then substitute these coefficients into (9) and obtain the following expression for $\tilde{f}_{l,m,n}$, the Fourier coefficients of $f(\mathbf{r})$:

$$\tilde{f}_{l,m,n} = \frac{1}{L_x L_y L_z}\sum_{i=0}^{N}\sum_{j=0}^{N-i}\sum_{k=0}^{N-i-j}\hat{f}_{i,j,k}\tilde{\psi}_i\left(\frac{l}{L_x};\lambda\right)\tilde{\psi}_j\left(\frac{m}{L_y};\lambda\right)\tilde{\psi}_k\left(\frac{n}{L_z};\lambda\right). \quad (21)$$

These values can be computed in $O(M^3N + M^2N^2 + M^3N)$ steps (see Appendix *B*).

2.7. Implementation details and running time

We chose to demonstrate the potential of the Hermite basis by implementing the rigid-body fitting of an atomistic structure of a protein in an electron-density map of low resolution. The *HermiteFit* algorithm was implemented using the C++ programming language and compiled using g++ with -O3 optimization. The running times of the tested algorithms were measured on a single core of an Intel Xeon CPU X5650 @2.67 GHz processor with 24 GB of RAM on a Linux 64-bit operating system.

Our fitting method typically samples some 10^{10} rigid-body configurations. Therefore, it is practical to group its fitting solutions into clusters. There are multiple ways to measure the similarity between rigid-body solutions. For example, the pairwise root-mean-square deviation (r.m.s.d.) is a fast and well accepted similarity measure. Thus, we clustered the fitting solutions using the rigid-body clustering algorithm implemented with the RigidRMSD library (Popov & Grudinin, 2014) as follows. Firstly, the fitting solution with the best score (yet unassigned to any cluster) is taken as the seed for the new cluster. Secondly, the pairwise r.m.s.d.s between the seed and all other predictions are measured and predictions with an r.m.s.d. lower than a certain threshold are put into the cluster. Finally, these two steps are iterated until all fitting predictions are assigned to corresponding clusters.

3. Analysis

This section provides analytical and numerical analysis of the density encoding in the Hermite basis. More specifically, we provide the choice of optimal model parameters and assess the quality of encoding.

3.1. Choice of parameters of the method

Orthogonal Hermite functions (6) decay exponentially after a certain distance and thus can encode information only within some interval. We can estimate this interval using the formula for the last root of a Hermite polynomial, $\xi_{1,N} \simeq (1+2N)^{1/2}/\lambda$ (Ricci, 1995), which gives an approximation for the half-size of the bounding box that we can successfully encode:

$$L_{\mathrm{box}}/2 \lesssim \frac{(1+2N)^{1/2}}{\lambda}. \quad (22)$$

On the other hand, orthogonal Hermite functions are the eigenfunctions of the continuous Fourier transform (20).

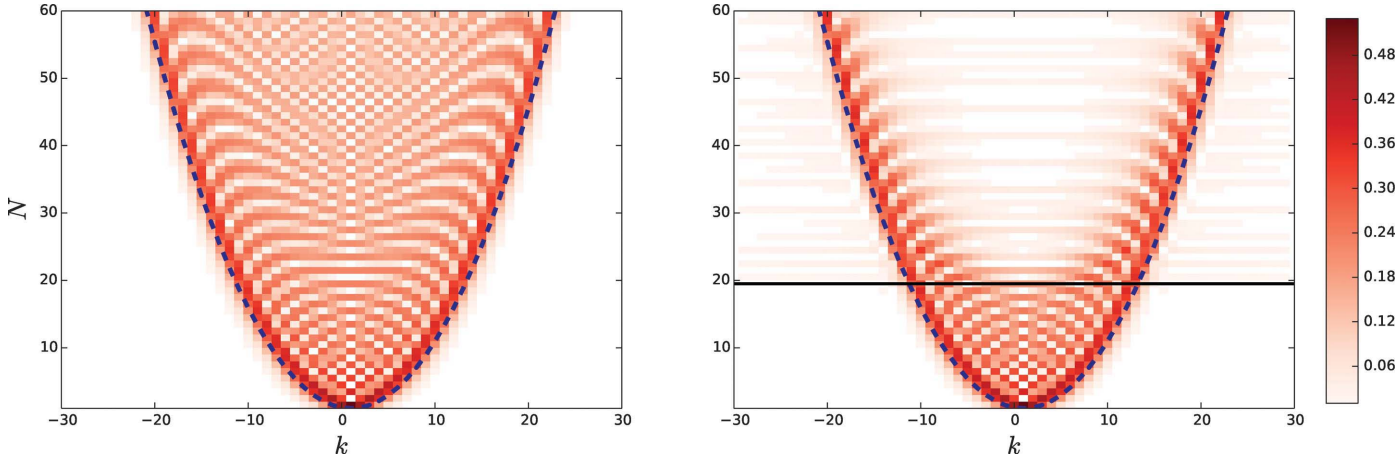


Figure 5
Absolute values of the two matrices $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$ that give the transfer matrix as their product (35). These matrices are computed with scaling parameter λ = 0.55 and input box size L_{box} = 23.0 Å, which mimics the first fitting example shown below. Left, $\mathbf{F}^{(1)}$, the scaled Fourier transform of a one-dimensional Hermite function, as given by (36). Right, $\mathbf{F}^{(2)}$, the scaled Fourier series of a one-dimensional Hermite function, as given by (37). The dashed blue line highlights the maximum encoded frequency according to (23). The solid black line in the right plot shows the maximum Hermite decomposition order N_{max} at which the two matrices are still identical (22).

Therefore, a Hermite decomposition of order N can encode only a certain interval of frequencies. Using the same approximation as in the case of the real-space interval, we obtain the following equation for the maximum encoding frequency:

$$\omega_{\max} = \frac{\lambda}{2\pi}(2N+1)^{1/2}. \tag{23}$$

In the case of the Fourier series expansion in the interval $(0, L_{\rm box})$, we can use the same estimation for the maximum encoding index $M_{\max}$ by setting $M_{\max} = 2L_{\rm box}\omega_{\max}$. The resolution ε of an X-ray electron-density map is defined by the size of the reciprocal lattice as $\varepsilon = 1/(2\omega_{\max})$ or, equivalently, $\varepsilon = L_{\rm box}/M_{\max}$. Therefore, using the resolution of the map ε and the order of the Fourier series expansion M, we can estimate the lower bound on the Hermite scaling parameter λ required

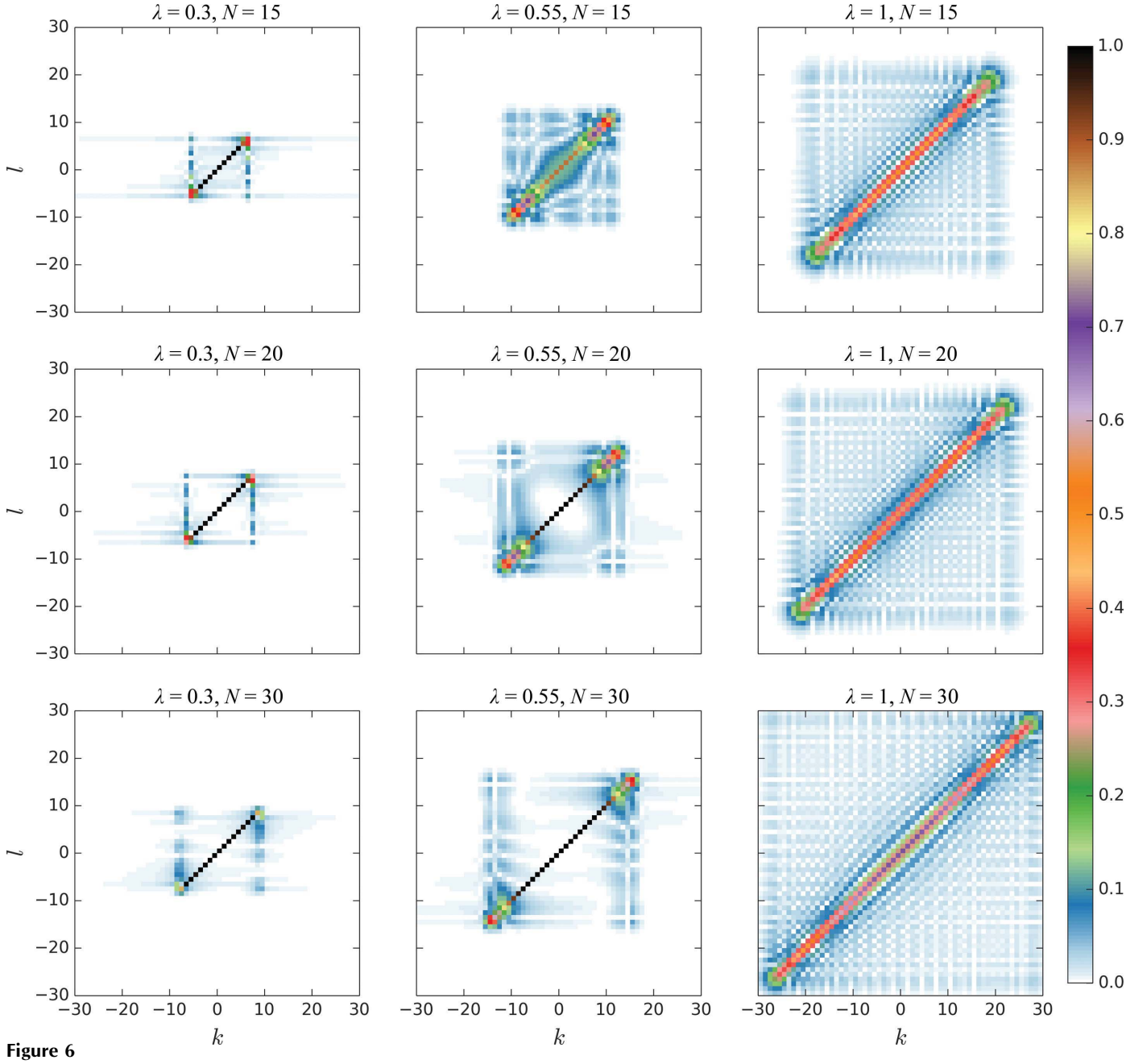


Figure 6
Nine examples of the absolute values of the transfer T matrices for three different values of λ and three different values of the Hermite decomposition order N. The number of Fourier coefficients is $M = 60$, and the input box size is $L_{\rm box} = 23.0$ Å, which mimics the first fitting example shown below. The Hermite decomposition orders are $N \in \{15, 20, 30\}$ and the parameter λ takes values of 0.3, 0.55 and 1.0 Å^{-1}. The first column corresponds to a relative $\lambda L_{\rm box}$ value of 6.9, the middle column corresponds to a relative $\lambda L_{\rm box}$ value of 12.6 and the right column to a relative $\lambda L_{\rm box}$ value of 23. Notably, at low values of λ the transfer matrix encodes only small-order reflections. The index of the last reflex can be estimated from (23) as $k_{\max} = (2N+1)^{1/2}\lambda L_{\rm box}/2\pi$. On increasing the value of λ, the number of encoded frequencies rises. However, at the same time the quality of the encoding of low frequencies worsens, as can be seen from the values on the diagonal.

to encode all the reflections of the electron-density diffraction pattern to be

$$\lambda \gtrsim \frac{\pi}{\max(\varepsilon, L_{\rm box}/M)(2N+1)^{1/2}}. \quad (24)$$

Here, we bounded the actual resolution by $L_{\rm box}/M$, because this will be the limit allowed by the finite Fourier series of order M.

The two inequalities (22) and (24) give approximate bounds on the scaling parameter λ, provided that we know the size of the box $L_{\rm box}$ containing protein density and the resolution of the map ε. Using these inequalities, we obtain the following relationship between the parameters λ and N:

$$\frac{\pi}{(2N+1)^{1/2}\max(\varepsilon, L_{\rm box}/M)} \lesssim \lambda \lesssim 2\frac{(1+2N)^{1/2}}{L_{\rm box}}, \quad (25)$$

which is valid for sufficiently large values of N. Nonetheless, we can use the following empirical estimation for the optimal value of λ at any N:

$$\lambda_{\rm opt} \simeq \frac{\pi}{2\max(\varepsilon, L_{\rm box}/M)(2N+1)^{1/2}} + \frac{(1+2N)^{1/2}}{L_{\rm box}}. \quad (26)$$

Using dimensionless relative parameters $\lambda L_{\rm box}$ and $L_{\rm box}/\varepsilon$, we may rewrite the previous expression as

$$\lambda L_{\rm box} \simeq \frac{\pi \min(L_{\rm box}/\varepsilon, M)}{2(2N+1)^{1/2}} + (1+2N)^{1/2}. \quad (27)$$

If at a given expansion order N there is no such parameter λ that satisfies inequality (25), then the protein representation might involve information loss. Therefore, we can estimate the minimum order $N_{\rm min}$ of the Hermite expansion that allows this inequality to have solutions to be

$$N_{\rm min} \simeq \frac{\pi}{4}\min\left(\frac{L_{\rm box}}{\varepsilon}, M\right). \quad (28)$$

The validity of the provided estimates and the graphical representation of the real-space and reciprocal-space bounds on the parameter λ will be demonstrated in the following sections.

The maximum order of the Fourier expansion $M_{\rm max}$ can be estimated from the resolution and the size of the density map as $\varepsilon = L_{\rm box}/M_{\rm max}$. However, when finding the global maximum of the cross-correlation function, we need to sample the space of possible translations of a protein with respect to the EDM with a step several times finer than the EDM resolution ε. In protein crystallography, it is common practice to set the sampling step size to $\varepsilon/3$ (Afonine & Urzhumtsev, 2004). In principle, we can use the same reasoning in choosing the optimal number of rotations $N_{\rm rot}$. When using spherical harmonics, the angular search step usually equals the resolution of the basis, $2\pi/N$ (Garzón *et al.*, 2007). In the case of the Hermite basis, we propose use of the same criterion.

3.2. The transfer matrix

Below, we describe an analytical model of encoding by the Hermite basis for the one-dimensional case. Suppose we have a function $f(x)$ that describes the electron density of a nonperiodic object. Without loss of generality, we assume that this function is defined in a one-dimensional interval of $(-L_{\rm box}/2;\ +L_{\rm box}/2)$. This function has the following decomposition into Fourier series:

$$\tilde{f}_k^{\rm exact} = \frac{1}{L_{\rm box}}\int_{-L_{\rm box}/2}^{+L_{\rm box}/2} f(x)\exp(-2\pi i k x/L_{\rm box})\,{\rm d}x. \quad (29)$$

We will refer to Fourier coefficients obtained using this expression as *exact*. The original function is then recovered by the inverse Fourier transform:

$$f(x) = \sum_{k=-\infty}^{+\infty} \tilde{f}_k^{\rm exact}\exp(2\pi i k x/L_{\rm box}). \quad (30)$$

On the other hand, our algorithm computes *approximate* Fourier coefficients using the Hermite-to-Fourier transform:

$$\tilde{f}_k^{\rm approx} = \frac{1}{L_{\rm box}}\sum_{n=0}^{N}\hat{f}_n\tilde{\psi}_n\left(\frac{k}{L_{box}};\lambda\right). \quad (31)$$

Assuming that the function $f(x)$ is zero outside the bounding interval, Hermite coefficients $\hat{f}_n$ can be written as the finite integral

$$\hat{f}_n = \int_{-L_{\rm box}/2}^{+L_{\rm box}/2} f(x)\psi_n(x;\lambda)\,{\rm d}x. \quad (32)$$

Now, we can express the approximate Fourier coefficients as a linear combination of the exact ones:

$$\tilde{f}_k^{\rm approx} = \sum_{l=-\infty}^{+\infty} T_{k,l}\tilde{f}_l^{\rm exact}, \quad (33)$$

where the transfer matrix $T_{k,l}$ is given by

$$T_{k,l} = \frac{1}{L_{\rm box}}\sum_{n=0}^{N}\tilde{\psi}_n(k;\lambda)\int_{-L_{\rm box}/2}^{+L_{\rm box}/2}\psi_n(x;\lambda)\exp(2\pi i l x/L_{\rm box})\,{\rm d}x. \quad (34)$$

The transfer matrix acts as a linear filter in reciprocal space and demonstrates how the input function is distorted by the finite size N of the Hermite basis. We should note that, generally, its values are complex numbers.

This matrix can also be seen as a product of two matrices,

$$\mathbf{T} = \mathbf{F}^{(1)}\mathbf{F}^{(2)}, \quad (35)$$

where the first matrix is a scaled Fourier transform of the basis functions,

$$F_{kn}^{(1)} = \tilde{\psi}_n(k;\lambda)/(L_{\rm box})^{1/2}, \quad (36)$$

and the second matrix is a scaled Fourier series of the basis functions,

$$F_{nl}^{(2)} = \int_{-L_{\rm box}/2}^{+L_{\rm box}/2}\psi_n(x;\lambda)\exp(2\pi i l x/L_{\rm box})\,{\rm d}x/(L_{\rm box})^{1/2}. \quad (37)$$

Fig. 5 shows the absolute values of the matrices $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$ computed with $\lambda = 0.55$ and $L_{\rm box} = 23$ Å. The values of the Fourier series $\mathbf{F}^{(2)}$ were computed numerically using adaptive quadrature. The dashed blue line shows the maximum

encoding frequency ω_{max}, according to (23), and bounds the encoding region. The solid black line on the right plot demonstrates the maximum order of the Hermite expansion (22), after which the Fourier series mainly encodes the frequencies near ω_{max}. This is because in a finite interval $(-L_{box}/2, +L_{box}/2)$ high-order Hermite basis functions become orthogonal to low-order Fourier basis functions.

Fig. 6 shows several examples of the absolute values of the transfer-matrix components for three different values of the Hermite scaling parameter λ and three values of the Hermite decomposition order N. The size of the transfer matrix was limited to 60 × 60 and the box size L_{box} was set to 23 Å. The ideal transfer matrix should be identity, which is only the case at $N\rightarrow\infty$, as we demonstrate below. We see, however, that the transfer matrix at small values of λ encodes only low-order reflections. The index of the last encoded reflex can be estimated from (23) as $k_{max} = (2N + 1)^{1/2}\lambda L_{box}/(2\pi)$. With the increase in order N and parameter λ, the number of encoded frequencies rises. At the same time, increasing the scaling parameter λ makes the quality of encoding of all of the frequencies worse, as we see in the right column. Therefore, it is very important to tune the value of λ according to the class of input functions, such that the quality of encoding becomes optimal. Below, we will assess encoding quality by means of the crystallographic R factor.

3.3. Asymptotic behaviour of the transfer matrix

Here, we demonstrate that the transfer matrix asymptotically achieves the Kronecker delta function at $N\rightarrow\infty$. Recall Mehler's formula (Mehler, 1866):

$$\sum_{n=0}^{N} u^n \psi_n(x)\psi_n(y) = \frac{1}{[\pi(1-u^2)]^{1/2}} \times \exp\left[-\frac{1-u}{1+u}\frac{(x+y)^2}{4} - \frac{1+u}{1-u}\frac{(x-y)^2}{4}\right]. \quad (38)$$

If we rewrite the transfer matrix in the following way,

$$T_{k,l} = \frac{1}{L_{box}}\int_{-L_{box}/2}^{+L_{box}/2} \mathrm{d}x \sum_{n=0}^{N}(-i)^n \psi_n\left(\frac{k}{L_{box}}; \frac{2\pi}{\lambda}\right) \times \exp(2\pi i l x/L_{box})\psi_n(x;\lambda), \quad (39)$$

and use the fact that

$$\psi_n(x;\lambda) \equiv \lambda^{1/2}\psi_n(\lambda x), \quad (40)$$

we see that we can use Mehler's formula to compute the limit

$$\lim_{N\rightarrow\infty}\sum_{n=0}^{N}(-i)^n \psi_n\left(\frac{k}{L_{box}}; \frac{2\pi}{\lambda}\right)\psi_n\left(\frac{l}{L_{box}};\lambda\right). \quad (41)$$

After a simple derivation, we obtain the final result,

$$\lim_{N\rightarrow\infty} T_{k,l} = \frac{1}{L_{box}}\int_{-L_{box}/2}^{+L_{box}/2} \exp(2\pi i l x/L_{box})\exp(-2\pi i k x/L_{box})\,\mathrm{d}x, \quad (42)$$

which is exactly the Kronecker delta function.

3.4. Encoding quality

There are several ways to evaluate the quality of a model encoding with the subsequent reconstruction. For example, in the optimal control theory (Boyd, 1991), the quality of a linear filter is estimated using a certain norm of the transfer matrix. However, in crystallography the most used quality criterion is the crystallographic R factor (Stout & Jensen, 1968),

$$R = \frac{\sum_l \left||\tilde{F}_l^{exact}| - |\tilde{F}_l^{mod}|\right|}{\sum_l |\tilde{F}_l^{exact}|}, \quad (43)$$

where F^{exact} and F^{mod} are the exact Fourier coefficients of a molecule and the coefficients computed from the Hermite coefficients, respectively. This quantity is a widely used measure of the agreement between a crystallographic model and the corresponding experimental X-ray diffraction data. In

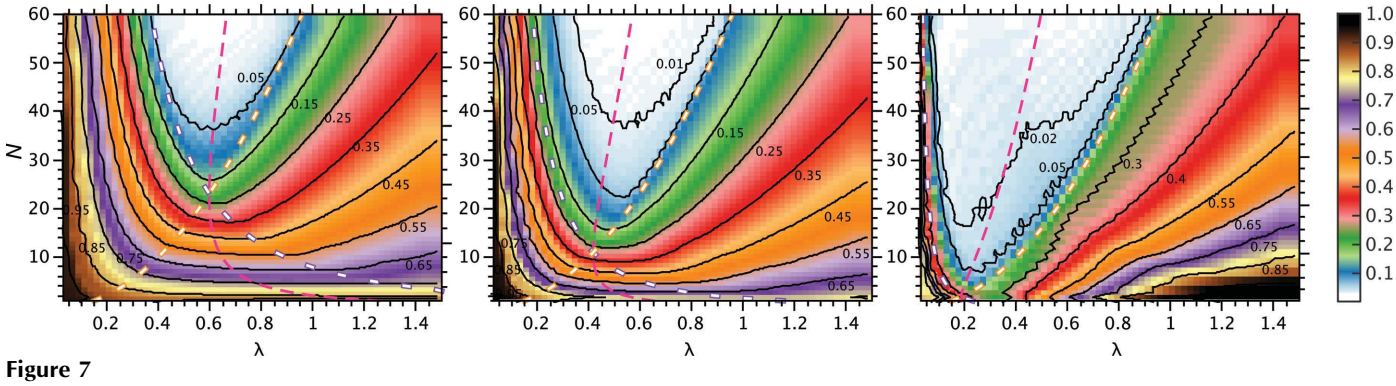


Figure 7
Analytical R factors in one dimension as a function of Hermite decomposition order N and scaling parameter λ computed at three different resolutions. The input signal is modelled as a sum of Gaussians (10) with a variance of $\alpha/2^{1/2}$ equispaced at a distance α. The number of Fourier coefficients is $M = 30$ and the input box size is $L_{box} = 23.0$ Å. These values were chosen to mimic the 1akg peptide decomposition. The estimate of the optimal parameter λ (26) is plotted as a red dashed line. The real-space bound on the optimal parameter λ (22) is shown as an orange dashed line. The reciprocal-space bound on the optimal parameter λ (24) is shown as a blue dashed line. Left, the Gaussian parameter $\alpha = 0.2$ Å, corresponding to an absolute input signal resolution of $\varepsilon = 0.31$ Å and a relative resolution $\varepsilon/L_{box} = 0.014$. However, in this case the actual absolute resolution is cut at $L_{box}/M = 0.77$ Å, which corresponds to a relative resolution of 0.033. Middle, the Gaussian parameter $\alpha = 1.0$ Å, corresponding to an absolute input signal resolution of $\varepsilon = 1.57$ Å and a relative resolution $\varepsilon/L_{box} = 0.068$. Right, the Gaussian parameter $\alpha = 5.0$ Å, corresponding to an absolute input signal resolution of $\varepsilon = 7.85$ Å and a relative resolution $\varepsilon/L_{box} = 0.34$.

the case of an ideal electron-density encoding, the R factor is equal to zero. In protein crystallography, models with R factors of less than 0.2 are regarded as good when working at a medium resolution.

The equations for the transfer matrix allow estimation of the R factor for certain classes of electron-density distributions. As described above (10), we use the Gaussian distribution to model the electron density of an atom. Exact Fourier coefficients of a molecule with N_{atoms} atoms at positions $\mathbf{r}_i$ are then given as

$$\tilde{f}^{\text{exact}}_{l,m,n}(\mathbf{s}) = \alpha^3\pi^{\frac{3}{2}} \sum_{i=1}^{N_{\text{atoms}}} \exp(-\alpha^2\pi^2\mathbf{s}^2_{lmn}) \exp(-2i\pi\mathbf{r}_i\mathbf{s}_{lmn}), \quad (44)$$

where $s_{l,m,n}$ is the wavevector and $\mathbf{s}_{l,m,n} = (l/L_x, m/L_y, n/L_z)$, where L_x, L_y and L_z are the dimensions of the bounding box along the corresponding axes. Similarly, one-dimensional exact Fourier coefficients of the Gaussian function are given as

$$\tilde{f}^{\text{exact}}_{l} = \alpha\pi^{\frac{1}{2}} \sum_{i=1}^{N_{\text{atoms}}} \exp(-\alpha^2 l^2/L^2_{\text{box}}) \exp(-2i\pi r_i l/L_{\text{box}}). \quad (45)$$

To see how the Hermite basis encodes Gaussian densities with various level of detail, we built models of the electron-density map with different parameters α. The width of the Gaussian determines the resolution of the density map according to

$$\varepsilon = \frac{\pi\alpha}{2}. \quad (46)$$

The derivation of this formula follows the one well known in crystallography which describes the extinction of diffraction reflections. For the sake of completeness, we provide its derivation in Appendix *C*.

To estimate the R factor for certain model parameters, we assume that the input electron density is given as a sum of Gaussians with variance of $\alpha/2^{1/2}$ equispaced at a distance α. Fig. 7 shows analytical R factors in one dimension computed using (33) and (45) as a function of the Hermite decomposition order N and the scaling parameter λ. We bounded the input and output frequencies by $M = 30$ Fourier coefficients. The size of the input interval L_{box} is set to 23.0 Å to mimic the α-conotoxin PnIB peptide (PDB entry 1akg) decomposition used in the fitting example below. We should stress that owing to the properties of the Hermite functions, the whole model is scale-invariant. More precisely, if we keep the product λL_{box} constant then the relative shape of the Hermite basis functions would not change. Also, if we scale L_{box} and α simultaneously then the value of the R factor is unchanged. Therefore, it is useful to provide relative resolutions computed as $\varepsilon/L_{\text{box}}$. Fig. 7 (left) shows R factors for the Gaussian parameter α = 0.2 Å corresponding to an absolute input signal resolution of ε = 0.31 Å and a relative resolution $\varepsilon/L_{\text{box}} = 0.014$. However, in this case the actual absolute resolution is cut at L_{box}/M = 0.77 Å, which corresponds to a relative resolution of 0.033. Fig. 7 (middle) shows R factors computed using the Gaussian parameter α = 1.0 Å corresponding to an absolute input signal resolution of ε = 1.57 Å and a relative resolution $\varepsilon/L_{\text{box}}$ = 0.068. Fig. 7 (right) shows R factors computed using the Gaussian parameter α = 5.0 Å corresponding to an absolute input signal resolution of ε = 7.85 Å and a relative resolution $\varepsilon/L_{\text{box}} = 0.34$. The estimate of the optimal parameter λ (26) is plotted as a red dashed line. The real-space bound on the optimal parameter λ (22) is shown as an orange dashed line. The reciprocal-space bound on the optimal parameter λ (24) is shown as a blue dashed line. We see that on lowering the resolution of the input signal the R factors decrease, as would be expected from general considerations. We can also see that the lower (22) and the upper (24) bounds on the optimal scaling parameter λ follow the isolines of the R-factor map. Therefore, their mean given by (22) provides a reasonable estimation of the optimal value of λ.

Fig. 8 shows R factors as a function of input signal resolution ε for three different Hermite decomposition orders N: 15, 20 and 30. R factors were estimated in the same way as in the previous case. More precisely, we assumed the same shape of the input electron density and then used (33) and (45) to compute the analytical R factors. For these plots, we computed the optimal scaling parameter λ using (22). The parameter L_{box} and the size of the transfer matrix M were constant and were equal to 23 Å and 30, respectively. As in the previous figure, these values are chosen to mimic the α-conotoxin PnIB peptide decomposition used in the fitting example below. The scale of the top horizontal axis gives the absolute resolution for L_{box} = 23 Å. The scale of the bottom horizontal axis gives the relative resolution. In order to compute the absolute resolution, its values need to be multiplied by the chosen value of L_{box}. As expected, the values of the R factors diminish as the resolution becomes lower. This is because at low resolutions low-frequency columns of the transfer matrix become more important. In the limiting cases of zero and infinite resolutions, the R factor can be computed directly from the transfer matrix as a certain norm of $T - I$. For the infinite resolution limit, it is given as the L_1 norm of the central

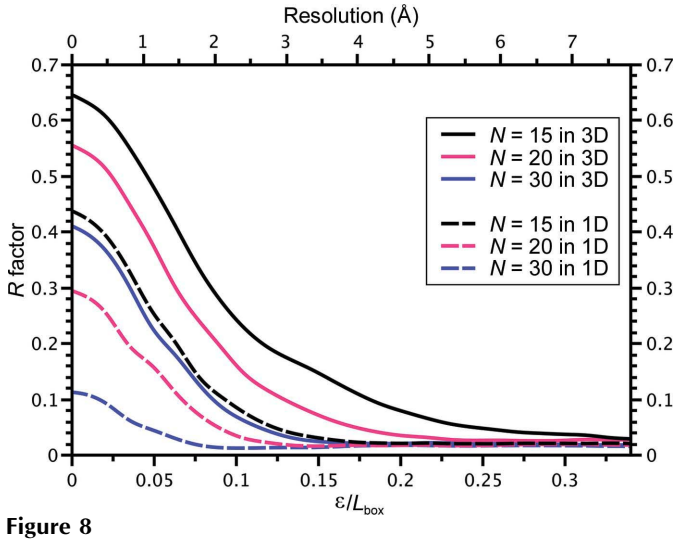


Figure 8
Analytical R factors in one and three dimensions as a function of the relative resolution $\varepsilon/L_{\text{box}}$. The absolute resolution at box size L_{box} = 23 Å is shown on the top horizontal axis. Plots for three different Hermite expansions orders are shown; $N \in \{15, 20, 30\}$. The parameters L_{box} and M were constant and were 23 Å and 30, correspondingly. The scaling parameter λ was estimated using (26).

column of the matrix $T - I$. For the zero resolution limit, the R factor is given by the entry-wise L_1 norm of $T - I$, $R = \sum_{i,j} |T_{i,j} - \delta_{i,j}|$. Fig. 8 also shows an estimation of R factors for the three-dimensional case. It is based on the assumption that Hermite decomposition encoding in three dimensions behaves similarly to the one-dimensional case, with the number of coefficients scaled as $N_{1D} = N_{3D}^{1/3}$.

4. Results and discussion

We tested and verified our algorithm using two examples of different difficulty. The first example is the small polypeptide α-conotoxin PnIB. We generated the EDM for this example from the coordinates of the polypeptide. The second example is the fitting the GroEL domains into the electron-density map of the GroEL complex.

4.1. α-Conotoxin PnIB

Firstly, we explored the relationship between the encoding quality and the quality of the fitting. For this purpose, we chose the small 16-residue polypeptide α-conotoxin PnIB. We downloaded the X-ray crystal structure of α-conotoxin PnIB (PDB code 1akg; Hu *et al.*, 1997) from the PDB (Berman *et al.*, 2000) and simulated the electron-density map ($2mF_o - DF_c$) using the Uppsala electron-density server (Kleywegt *et al.*, 2004) with resolution ε = 1.1 Å. We computed the protein density according to (10) with the Gaussian width α = 1.0 Å using only the non-H atoms of the standard amino acids. We rotated the initial 1akg structure by arbitrarily chosen Euler angles of $\varphi = 76^\circ$, $\theta = 234^\circ$ and $\psi = 56^\circ$ and used it as the input for the fitting workflow. We used N_{rot} = 500 (corresponding to an angular step of 36°) rotations represented with uniformly distributed Euler angles spanning the space $2\pi \times \pi \times 2\pi$. The order of the Hermite expansion was set to N = 15, which is the minimum expansion order allowed at this resolution according to (28). The order of the Fourier expansion was twice the order of the Hermite expansion: M = 30 for each dimension.

To see how the encoding quality influences the fitting algorithm, we studied the dependence of the decomposition on the scaling parameter λ. We chose a range of λ parameters between 0.05 and 2.0. For each λ, we computed the best fitting score along with the average fitting score. Fitting results are shown in Fig. 9. We see that by choosing a small λ we neglect the details of the protein structure (Fig. 9*a*) and therefore we cannot discriminate between different orientations of the protein (the maximum score for λ = 0.05 is very close to the average score). When choosing a sufficiently large λ, we obtain satisfactory discriminative power to find the near-native position of the protein (Figs. 9*c* and 9*d*). We also see that, for example for λ = 0.5, the difference between the maximum and the average score is much larger than in the case of λ = 0.05. Also, when we take too large a λ we cannot encode the whole protein (Fig. 9*e*). The red dashed line in Fig. 9 shows R factors computed with (43). We see that the choice of the parameter λ influences the R factors and thus determines the quality of the

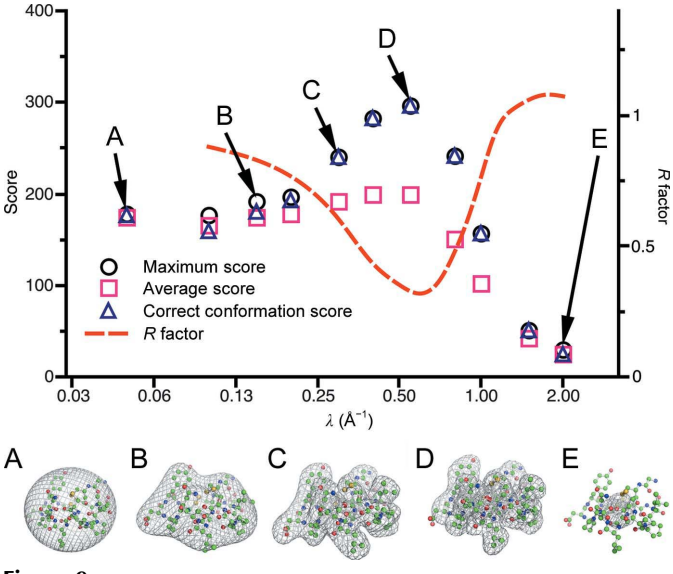


Figure 9
Test of the fitting algorithm on an artificially generated EDM for the α-conotoxin PnIB (PDB entry 1akg). Here, we plotted the dependence of four parameters, the maximum score, the average score, the score of the near-native conformation and the crystallographic R factor, on the scaling parameter λ. The isosurface of the Hermite decomposition at protein model density equal to $(\rho_{max} + \rho_{min})/2$ and several values of λ are shown in subplots A (λ = 0.05), B (λ = 0.15), C (λ = 0.3), D (λ = 0.55) and E (λ = 2.0).

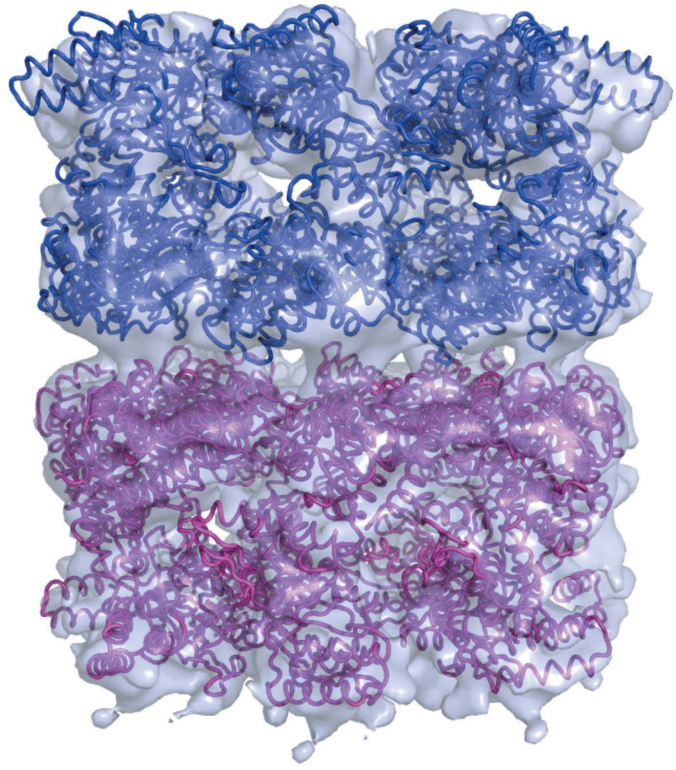


Figure 10
Result of fitting chain *A* of the GroEL–GroES X-ray structure (PDB entry 1aon) to the GroEL complex electron-density map (EMD-2001). Two heptameric rings are shown in different colours. The average r.m.s.d. measured using the C^α atoms between the two closest chains in the fitted structure and the flexibly refined structure provided by the authors of the EDM (PDB entry 4aau) is 5.35 Å.

fitting. Notably, the minimum of the *R*-factor curve corresponds to the maximum of the fitting score.

Owing to the strong influence of the scaling parameter λ on the discriminating power of the algorithm, we estimated its optimal value to gain the maximum separation between the score of the correct pose and the average score. Provided that the box that contains all of the rotations of the peptide has the size L_{box} = 23 Å and setting the resolution of the EDM ε = 1.1 Å, (26) gives an estimate of the optimal value of the scaling parameter: $\lambda_{opt} \simeq 0.50$. Fig. 9 shows that this estimation corresponds to the best discrimination between the near-native and all other structures, which can be deduced from the maximum separation between the score of the prediction and the average score. The r.m.s.d. between the prediction and the solution at this value of λ is 1.03 Å. We should note that the r.m.s.d can be decreased by taking a finer angular search step.

4.2. GroEL complex

Here, we demonstrate that our approach gives essentially the same results as other programs, provided that the scoring function is the same (LCCF in this case). For this purpose, we use a classical test for a fitting algorithm: the GroEL complex map. We downloaded the EDM of the GroEL complex from the Electron Microscopy Data Bank (EMDB; code EMD-2001) with a resolution of 8.5 Å. We then downloaded the crystal structure of the GroEL subunits from the PDB database. We used the GroEL–GroES complex structure (PDB entry 1aon; Xu *et al.*, 1997), from which we extracted chain *A*, centred it and arbitrarily rotated it to exclude any bias. We chose the sampling grid size according to the resolution and the size of the EDM. The EDM was first padded with zeros and then transformed to the Fourier basis using the *FFT* algorithm. The number of coefficients in the Fourier decomposition *M* was equal to 105 × 107 × 119. The angular search step was set to 30°. We used a Hermite expansion order of *N* = 15, which is larger than the minimum expansion order allowed at this resolution, $N_{min} \simeq 9$ (see equation 28). We sampled the rotations using the spiral algorithm (Saff & Kuijlaars, 1997), which generates an equispaced distribution of points on a sphere. Unlike in the previous example, owing to the lower resolution of the GroEL EDM, here we fitted Laplacian-filtered protein density into the Laplacian-filtered EDM.

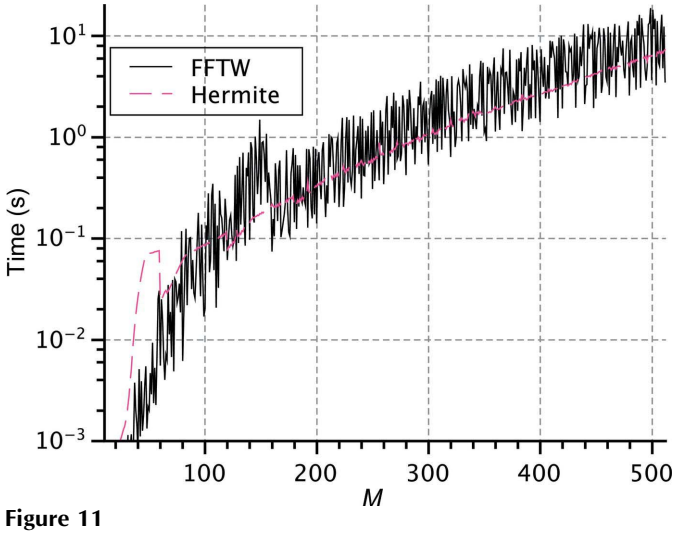


Figure 11
Running times of the Hermite-to-Fourier space transition performed using our algorithm and the *FFT* algorithm on a cubic grid of $M \times M \times M$ as a function of the Fourier expansion order *M*. We used the FFTW3 library (Frigo & Johnson, 2005) with the double-precision real discrete Fourier transform using the flag FFTW_ESTIMATE to measure the speed of the FFT. The order of the Hermite expansion was *N* = 15.

Table 2
Comparison of the *HermiteFit* algorithm with the *colores* and *ADP_EM* algorithms.

The comparison criteria were chosen to be the total running time and the running time per point of search space

Algorithm	No. of rotation-space points	No. of translation-space points	Runtime (s)	Time per point ($\times 10^{-7}$ s)
ADP_EM	16384	23186	139	3.6
colores	4416	1336965	1454	2.5
HermiteFit	4416	1336965	917	1.5

Table 3
Comparison of the models obtained using the *HermiteFit*, *colores* and *ADP_EM* algorithms with the model obtained by the authors of the electron-density map (PDB entry 4aau).

For each pair of models, the r.m.s.d. was measured using the C^{α} atoms and the centres of mass of the corresponding chains and was then averaged over all of the chains comprising the assembly.

Algorithm	R.m.s.d., C^{α} (Å)	R.m.s.d., centres of mass (Å)
ADP_EM	4.61	2.29
colores	5.42	2.52
HermiteFit	5.35	2.64

After the six-dimensional exhaustive search, we clustered the solutions using a clustering threshold of 10 Å and kept the top 14 poses. All 14 poses corresponded to individual chains of the complex, which is comprised of two heptameric rings. Fig. 10 shows the results of the fitting. We compared the fitted model with the model provided by the authors of the EDM (PDB entry 4aau; Clare *et al.*, 2012). The average r.m.s.d. between the chains owing to the flexible deformations measured using C^{α} atoms was 3.0 Å. More precisely, we superposed the corresponding chains of both models using rigid-body transformations and then measured the r.m.s.d. between them. Overall, the average r.m.s.d. between C^{α} atoms was 5.35 Å. This includes both the discrepancy between corresponding chains in the assembly arising from flexible deformations and the rigid-body misfit. The average distance between the centres of mass of the corresponding chains was 2.64 Å (Table 3).

4.3. Runtime of Hermite- to Fourier-space transition

The use of the fast Fourier transform has been an inevitable step in every fitting algorithm until now. Instead, we introduced a basis from which we can transform a decomposition into the Fourier basis avoiding evaluation of the FFT on a grid. When the grid becomes large, the asymptotic complexity of our algorithm becomes $O(M^3N)$ (see equation 21). It is comparable to the complexity of the fast Fourier transform

algorithm, $O(M^3 \log M)$. Intuitively, at large orders of the Fourier expansion M our algorithm should be faster compared with the FFT. Thus, we conducted a numerical experiment to compare the actual running times. Fig. 11 shows the time needed to compute the FFT on a cubic grid of size M and the time needed to transform a Hermite expansion of order $N = 15$ to the same Fourier grid. We can see that, generally, at large values of M, $M > \approx 100$, the transition from Hermite space into Fourier space is faster compared with the speed of the FFT. Also, the timing of the transition grows evenly with respect to M in contrast to the timing of the FFT. One has to take into account that we compared our algorithm with the highly optimized FFTW3 library (Frigo & Johnson, 2005). It is probable that additional optimization of *HermiteFit* could improve performance even further. One of the ways to speed up the transition will be to use the fast Hermite transform instead of naive matrix multiplication (Leibon *et al.*, 2008). This implementation will be the subject of future work.

4.4. Comparison with *Situs* and *ADP_EM*

We compared the *HermiteFit* algorithm with two popular existing fitting methods: the *colores* program from the *Situs* package (Chacón & Wriggers, 2002) and the *ADP_EM* fitting tool (Garzón *et al.*, 2007). These two packages represent the two major approaches to the problem of exhaustive search in six-dimensional space of rigid-body motions. *Colores*, a widely used CCF-based fitting tool, rapidly scans the translational degrees of freedom using the fast Fourier transform. The rotations, however, are sampled exhaustively by enumerating a list of equispaced distributed rotations on a sphere. *ADP_EM* choses points in real space, places the atomic structure there and then rotationally matches it to the EDM using the fast rotational matching algorithm. The authors of *ADP_EM* compared their algorithm with five-dimensional rotational matching and found that three-dimensional rotational matching works faster in practice (Garzón *et al.*, 2007).

For comparison, we normalized the running times of the fitting algorithms by the sizes of the search space. For *colores* and *HermiteFit* the size of the search space is equal to the number of grid cells (M^3 for a cubic grid in the *HermiteFit* algorithm) multiplied by the number of sampled angles. The size of the search space of the *ADP_EM* algorithm is the number of points in real space times the number of cells of the angular grid. The latter is built from uniformly sampled Euler angles on a grid of $2\pi \times \pi \times 2\pi$. The size of the angular grid is determined by the order $N_{\rm exp}$ of a spherical harmonics expansion and equals $4N_{\rm exp}^3$. For *colores* and *HermiteFit*, we used an angular search step of 30°. The resolution of the EDM for *colores* and *HermiteFit* was set to 8.5 Å. The Fourier grid that was used by *colores* and the *HermiteFit* algorithm had dimensions of $105 \times 107 \times 119$. For *ADP_EM*, we used a spherical harmonics expansion order of $N_{\rm exp} = 16$.

Table 2 shows the normalized times of the complete six-dimensional search for the three algorithms in the case of fitting the GroEL subunit into the 8.5 Å resolution GroEL electron-density map. Judging by the total running time, *ADP_EM* has a large advantage over the two other algorithms, which exhaustively search all of the space of possible translations. However, in terms of running time per search point, the *HermiteFit* algorithm is more effective than the other two. Interestingly, *colores* spends about half of the total search time on computation of the Fourier coefficients of the rotated protein. Therefore, it was very important for us to speed up this step. Nonetheless, all three tested algorithms have their own advantages and drawbacks. For example, *ADP_EM* can use smart heuristics to reduce the number of search points in real space. However, its sample points in the space of rigid-body rotations are distributed non-uniformly. In particular, rotations near the poles are sampled more densely, making this sampling scheme less effective (Saff & Kuijlaars, 1997). On the other hand, the *HermiteFit* algorithm along with the *colores* algorithm sample the rotational space nearly uniformly using the spiral algorithm while the translational space sampling also remains uniform. We would like to stress that the absolute runtimes (shown in Table 2) are not very informative. In particular, they dramatically depend on the choice of the FFT library, code optimization, the choice of compiler and compilation options *etc.* However, this comparison clearly demonstrates that the new approach paves the way to speed up one of the bottlenecks of fitting methods: the projection of the rotated structure into the Fourier space.

To assess the fitting quality of the tested methods, we measured the r.m.s.d.s between the obtained models and the structure obtained by the authors of the electron-density map (PDB entry 4aau). Table 3 shows a comparison of the measured r.m.s.d.s for *ADP_EM*, *colores* and *HermiteFit*. We used two different criteria for the measurements. Firstly, we measured the average r.m.s.d. between C^α atoms. Secondly, we measured the average distance between the centres of mass of the corresponding chains. *ADP_EM* produced a model with an r.m.s.d. of 4.61 Å from the solution; the r.m.s.d.s for *colores* and *HermiteFit* were 5.42 and 5.35 Å, respectively. Clearly, Table 3 demonstrates that the tested algorithms produce equal quality models. However, the results of *ADP_EM* are slightly better, presumably because of the finer rotational sampling.

5. Conclusion

In this paper, we have presented *HermiteFit*, a new method that performs an exhaustive search in the six-dimensional space of rigid-body motions. It uses orthogonal Hermite functions to encode the electron density and performs the critical steps of the fitting workflow in Hermite space. As part of the new method, we developed an algorithm for the rotation of the decomposition in the Hermite basis and an algorithm for the conversion of the Hermite expansion coefficients into the Fourier basis. By introducing the Hermite decomposition into the EDM fitting workflow, we inevitably introduced an additional scaling parameter λ. For this parameter, we provided tight bounds and an estimation of the optimal value that depends only on the properties of the fitting problem and the desired order of the polynomial decomposition (equations 25 and 26). Using two examples, we

demonstrated the validity of these bounds as well as the sufficiency of the Hermite expansion of order $N = 15$ to solve the standard EDM fitting problems. In particular, we derived a formula for Laplacian-filtered Hermite decomposition and employed this result to fit a single chain of GroEL into its electron-density map. Using analytical analysis, we calculated the crystallographic R factor produced by our method, which does not depend on the particular density that we encode. This allowed us to avoid tuning of fitting parameters and provided a clear understanding of the error sources in the algorithm. Finally, we compared our algorithm with two widely used fitting methods: *ADP_EM* and *colores* from the *Situs* package.

The proposed algorithm can be straightforwardly applied to other problems in structural bioinformatics such as, for example, protein–protein and protein–ligand docking. It can also be used for computer vision and three-dimensional object-recognition problems. The improvement in the speed of the algorithm may have an impact on flexible protein docking, flexible EDM fitting and other difficult problems that require a six-dimensional exhaustive space search as their initial step.

HermiteFit will be made available stand-alone at http://nano-d.inrialpes.fr/software/HermiteFit and as a plugin for the *SAMSON* software platform, and is available upon request from the authors.

APPENDIX *A*
Shifted Gaussian expansion

Here, we provide the derivation of the expansion coefficients of a shifted Gaussian of the form

$$g(\mathbf{r}) = \exp\left(-\frac{|\mathbf{r}-\mathbf{r}_0|^2}{\alpha^2}\right) \tag{47}$$

into the orthogonal Hermite basis. The well known property of this basis (as well as of any orthogonal basis) is the following:

$$\begin{gathered}\text{if } f(x,y,z) = f^{(1)}(x)f^{(2)}(y)f^{(3)}(z)\\ \text{and } f^{(k)}(t) = \sum_{i=0}^{N}\hat{f}_i^{(k)}\psi_i(t;\lambda)\\ \text{then}\\ \hat{f}_{i,j,k} = \hat{f}_i^{(1)}\hat{f}_j^{(2)}\hat{f}_k^{(3)}.\end{gathered} \tag{48}$$

Firstly, we derive the decomposition of a one-dimensional Gaussian into the one-dimensional orthogonal Hermite basis. Then, using property (48), we obtain the decomposition of a three-dimensional Gaussian into the three-dimensional orthogonal Hermite basis. More specifically, the one-dimensional Gaussian function is

$$g(x) = \exp\left[-\frac{(x-\xi)^2}{\alpha^2}\right]. \tag{49}$$

Its decomposition coefficients are

$$\begin{aligned}\hat{g}_n(\xi;\lambda,\alpha) &= \int g(x)\psi_n(x;\lambda)\,\mathrm{d}x\\ &= \frac{n!\lambda^{1/2}\exp\left[-\frac{\xi^2}{\alpha^2}\left(1-\frac{1}{\alpha^2\beta^2}\right)\right]}{\left(2^n n!\pi^{1/2}\right)^{1/2}}\sum_{m=0}^{(n/2)}\frac{(-1)^m}{m!(n-2m)!}\\ &\times\int\exp\left[-\beta^2\left(x-\frac{\xi}{\alpha^2\beta^2}\right)^2\right]\left[2\lambda\left(x-\frac{\xi}{\alpha^2\beta^2}\right)+\frac{2\lambda\xi}{\alpha^2\beta^2}\right]^{n-2m}\mathrm{d}x\end{aligned} \tag{50}$$

where $\beta^2 = (\lambda^2/2) + (1/\alpha^2)$. From now on, we will, for brevity, write $\hat{g}_n$ instead of $\hat{g}_n(\xi;\lambda,\alpha)$. Changing the variables $t = x - (\xi/\alpha^2\beta^2)$ and denoting $a = \xi/\alpha^2\beta^2$, we obtain

$$\begin{aligned}\hat{g}_n &= \frac{n!\lambda^{1/2}\exp\left[-\frac{\xi^2}{\alpha^2}\left(1-\frac{1}{\beta^2}\right)\right]}{\left(2^n n!\pi^{1/2}\right)^{1/2}}\sum_{m=0}^{(n/2)}\frac{(-1)^m(2\lambda)^{n-2m}}{m!(n-2m)!}\\ &\times\int\exp(-\beta^2t^2)(t+a)^{n-2m}\,\mathrm{d}x.\end{aligned} \tag{51}$$

Next, we decompose the sum $(t+a)^k$ using Newton's formula,

$$(t+a)^k = \sum_{i=0}^{k}\binom{k}{i}t^i a^{k-i}. \tag{52}$$

Thus, the integral in (51) will read

$$\begin{aligned}&\int\exp(-\beta^2t^2)(t+a)^{n-2m}\,\mathrm{d}x\\ &\quad=\sum_{i=0,i\text{-even}}^{n-2m}\frac{(n-2m)!}{2^i(i/2)!(n-2m-i)!}\pi^{1/2}\beta^{-1-i}a^{n-2m-i}.\end{aligned} \tag{53}$$

Substituting it into the formula for $\hat{g}_n$ and denoting $\sum_{i=0,i\text{-even}}^{n-2m} = \sum_{l=0}^{[(n-2m)/2]}$ $(i = 2l)$, we obtain the following expression for the coefficients,

$$\begin{aligned}\hat{g}_n(\xi;\lambda,\alpha) &= \exp\left[-\frac{\xi^2}{\alpha^2}\left(1-\frac{1}{\alpha^2\beta^2}\right)\right]\left(\frac{n!\pi^{1/2}\lambda}{2^n}\right)^{1/2}\\ &\times\sum_{m=0}^{(n/2)}\sum_{l=0}^{[(n-2m)/2]}\frac{(-1)^m2^{n-2m-2l}\lambda^{n-2m}}{l!(n-2m-2l)!m!}\beta^{-2n+4m+2l-1}\left(\frac{\xi}{\alpha^2}\right)^{n-2m-2l}.\end{aligned} \tag{54}$$

Finally, using (48) we obtain a decomposition of the three-dimensional Gaussian into the three-dimensional Hermite basis. We should note that in order to avoid the rounding error, one should begin the summation with the Gaussians that are located farther from the origin.

APPENDIX *B*
Fast summation

Here we explain the fast summation in (55),

$$\tilde{f}_{l,m,n} = \sum_{i=0}^{N}\sum_{j=0}^{N-j}\sum_{k=0}^{N-i-j}\hat{f}_{i,j,k}\tilde{\psi}_{i,l}\tilde{\psi}_{j,m}\tilde{\psi}_{k,n}, \tag{55}$$

with indices $l, m, n \in (0, M)$. The summation in this formula can be performed with less operations than a naive estimation $O(M^3N^3)$ suggests. We perform the fast summation by splitting the equation into three consecutive sums:

$$\widetilde{T^1_{i,j,n}} = \sum_{k=0}^{N-i-j}\hat{f}_{i,j,k}\tilde{\psi}_{k,n}, \tag{56}$$

$$\widetilde{T^2_{i,m,n}} = \sum_{j=0}^{N-i} \widetilde{T^1_{i,j,n}} \tilde{\psi}_{j,m}, \quad (57)$$

$$\tilde{f}_{l,m,n} = \sum_{i=0}^{N} \widetilde{T^2_{i,m,n}} \tilde{\psi}_{i,l}. \quad (58)$$

It is easy to see that the construction of the $\widetilde{T^1_{i,j,n}}$ matrix takes $O(MN^3)$ operations, the construction of the $\widetilde{T^2_{i,m,n}}$ matrix takes $O(M^2N^2)$ operations and the final summation takes $O(M^3N)$ operations. In the common use case ($N = 15$, $M >> N$) the last sum takes much more time than the other two. To optimize it, we used the Gaussian method to multiply complex numbers and expressed the whole sum as a generalized matrix product of three real-valued matrices. To implement these operations, we used the ATLAS library.

APPENDIX C
Resolution model

To illustrate the connection between the parameter α in the model of electron density (10) and the resolution of the X-ray diffraction pattern, we use the simplest model. More precisely, we model the electron density as the array of Gaussians in a perfect one-dimensional lattice perpendicular to the incoming radiation beam. The parameter α then plays the role similar to the temperature B factor. X-ray diffraction intensity depends on the angle between the incoming beam and the direction to the detector θ as

$$I \propto \left| \int f(x) \exp\left(2\pi i x \frac{\sin\theta}{\lambda}\right) \mathrm{d}x \right|^2, \quad (59)$$

where λ is the wavelength of the incoming radiation. Using the model density (10), we obtain

$$I \propto \left| \alpha\pi^{1/2} \exp\left[-\left(\pi \frac{\sin\theta}{\lambda}\alpha\right)^2 \right] \int \rho(x) \exp\left(2\pi i x \frac{\sin\theta}{\lambda}\right) \mathrm{d}x \right|^2, \quad (60)$$

where $\rho(x)$ is the sum of delta functions at the atomic positions. Therefore, the extinction of the diffraction peaks is proportional to $|\exp\{-[\pi(\sin\theta/\lambda)\alpha]^2\}|^2$, where we neglect the quadratic factor before the exponential.

According to the definition used in crystallography, resolution is the interplanar distance in real space corresponding to the last observable peak in reciprocal space. Unfortunately, the index of the last peak depends on the detector's noise and strongly depends on the characteristics of the measurement device. Therefore, to give a qualitative estimation of the dependence of resolution on the model parameter α, we assume that the last observable peak is that whose intensity decreases approximately by the factor e^2. The corresponding angle is then given by

$$\sin\theta_{\max} = \frac{\lambda}{\pi\alpha}. \quad (61)$$

Therefore, the minimum interplanar distance, or the resolution, is given by Bragg's law as

$$\varepsilon = \pi\alpha/2. \quad (62)$$

The authors wish to thank David W. Ritchie from Loria Nancy and Jorge Navaza from the Institute of Structural Biology, Grenoble for their advice, comments and suggestions during the preparation of the manuscript. The authors also wish to thank the journal co-editor Vladimir Y. Lunin and the two anonymous reviewers for numerous comments that helped to improve the manuscript. This work was supported by the Agence Nationale de la Recherche (ANR-11-MONU-006-01 and ANR-11-MONU-006-03) and the Centre National de la Recherche Scientifique (PEPS Bio-Math-Info 2012–2013).

References

Afonine, P. V. & Urzhumtsev, A. (2004). *Acta Cryst.* A**60**, 19–32.
Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. & Bourne, P. (2000). *Nucleic Acids Res.* **28**, 235–242.
Boyd, S. (1991). *Linear Controller Design: Limits of Performance*. Englewood Cliffs: Prentice Hall.
Chacón, P. & Wriggers, W. (2002). *J. Mol. Biol.* **317**, 375–384.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
Cheng, Y. & Walz, T. (2009). *Annu. Rev. Biochem.* **78**, 723–742.
Clare, D. K., Vasishtan, D., Stagg, S., Quispe, J., Farr, G. W., Topf, M., Horwich, A. L. & Saibil, H. R. (2012). *Cell*, **149**, 113–123.
Crowther, T. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–178. New York: Gordon & Breach.
Frigo, M. & Johnson, S. G. (2005). *Proc. IEEE*, **93**, 216–231.
Gabb, H., Jackson, R. & Sternberg, M. (1997). *J. Mol. Biol.* **272**, 106–120.
Garzón, J. I., Kovacs, J., Abagyan, R. & Chacón, P. (2007). *Bioinformatics*, **23**, 427–433.
Hu, S.-H., Gehrmann, J., Alewood, P. F., Craik, D. J. & Martin, J. L. (1997). *Biochemistry*, **36**, 11323–11330.
Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A., Aflalo, C. & Vakser, I. (1992). *Proc. Natl Acad. Sci. USA*, **89**, 2195–2199.
Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Kovacs, J. A., Chacón, P., Cong, Y., Metwally, E. & Wriggers, W. (2003). *Acta Cryst.* D**59**, 1371–1376.
Kovacs, J. A. & Wriggers, W. (2002). *Acta Cryst.* D**58**, 1282–1286.
Leibon, G., Rockmore, D. N., Park, W., Taintor, R. & Chirikjian, G. S. (2008). *Theor. Comput. Sci.* **409**, 211–228.
Mehler, F. G. (1866). *J. Reine Angew. Math.* **66**, 161–176.
Navaza, J. (2002). *Acta Cryst.* A**58**, 568–573.
Navaza, J. & Vernoslova, E. (1995). *Acta Cryst.* A**51**, 445–449.
Park, W., Leibon, G., Rockmore, D. & Chirikjian, G. (2009). *IEEE Trans. Image Process.* **18**, 1988–2003.
Popov, P. & Grudinin, S. (2014). *J. Comput. Chem.* **35**, 950–956.
Ricci, P. (1995). *Comput. Math. Appl.* **30**, 409–416.
Ritchie, D. W., Kozakov, D. & Vajda, S. (2008). *Bioinformatics*, **24**, 1865–1873.
Rossmann, M., Bernal, R. & Pletnev, S. (2001). *J. Struct. Biol.* **136**, 190–200.
Saff, E. B. & Kuijlaars, A. B. (1997). *Math. Intell.* **19**, 5–11.
Sayre, D. (1951). *Acta Cryst.* **4**, 362–367.
Siebert, X. & Navaza, J. (2009). *Acta Cryst.* D**65**, 651–658.
Stone, H. (1998). *IEEE Trans. Signal Process.* **46**, 2819–2821.
Stout, G. H. & Jensen, L. H. (1968). *X-ray Structure Determination: A Practical Guide*, Vol. 2. New York: Macmillan.
Suhre, K., Navaza, J. & Sanejouand, Y.-H. (2006). *Acta Cryst.* D**62**, 1098–1100.

Svergun, D. I. & Koch, M. H. J. (2003). *Rep. Prog. Phys.* **66**, 1735.
Ten Eyck, L. F. (1977). *Acta Cryst.* A**33**, 486–492.
Vasishtan, D. & Topf, M. (2011). *J. Struct. Biol.* **174**, 333–343.
Volkmann, N. & Hanein, D. (1999). *J. Struct. Biol.* **125**, 176–184.
Wriggers, W. (2010). *Biophys. Rev.*, **2**, 21–27.
Xu, Z., Horwich, A. L. & Sigler, P. B. (1997). *Nature (London)*, **388**, 741–750.