A Subgroup Algorithm to Identify Cross-Rotation Peaks Consistent with Non-Crystallographic Symmetry

Ryan H. Lilien
_Dartmouth College_

Chris Bailey-Kellogg
_Purdue University_

Amy C. Anderson
_Dartmouth College_

Bruce R. Donald
_Dartmouth College_

Follow this and additional works at: [https://digitalcommons.dartmouth.edu/cs_tr](https://digitalcommons.dartmouth.edu/cs_tr)

Part of the Computer Sciences Commons

**Dartmouth Digital Commons Citation**

This Technical Report is brought to you for free and open access by the Computer Science at Dartmouth Digital Commons. It has been accepted for inclusion in Computer Science Technical Reports by an authorized administrator of Dartmouth Digital Commons. For more information, please contact dartmouthdigitalcommons@groups.dartmouth.edu.
A Subgroup Algorithm to Identify Cross-Rotation Peaks Consistent with Non-Crystallographic Symmetry

Ryan H. Lilien*,†‡
Chris Bailey-Kellogg§
Amy C. Anderson†¶,**
Bruce R. Donald*,†¶,**

December 5, 2003

Abstract

Molecular replacement (MR) often plays a prominent role in determining initial phase angles for structure determination by X-ray crystallography. In this paper, an efficient quaternion-based algorithm is presented for analyzing peaks from a cross-rotation function to identify model orientations consistent with non-crystallographic symmetry (NCS), and to generate NCS-consistent orientations missing from the list of cross-rotation peaks. Our algorithm, crans, analyzes the rotation differences between each pair of cross-rotation peaks to identify finite subgroups of NCS. Sets of rotation differences satisfying the subgroup axioms correspond to orientations compatible with the correct NCS. The crans algorithm was first tested using cross-rotation peaks computed from structure factor data for three test systems, and then used to assist in the de novo structure determination of dihydrofolate reductase-thymidylate synthase (DHFR-TS) from Cryptosporidium hominis. In every case, the crans algorithm runs in seconds to identify orientations consistent with the observed NCS and to generate missing orientations not present in the cross-rotation peak list. The crans algorithm has application in every molecular replacement phasing effort with NCS.

Dartmouth Computer Science Department Technical Report No: TR2003-481
http://www.cs.dartmouth.edu/reports/reports.html

*Dartmouth Computer Science Department, Hanover, NH 03755, USA.
†Dartmouth Medical School, Hanover, NH 03755, USA.
‡Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755, USA.
§Purdue Department of Computer Sciences, West Lafayette, IN 47907, USA.
*†Dartmouth Department of Chemistry, Hanover, NH 03755, USA.
∥Dartmouth Department of Biological Sciences, Hanover, NH 03755, USA.
**Corresponding authors, Bruce.R.Donald@dartmouth.edu and Amy.C.Anderson@dartmouth.edu. This work is supported by grants to B.R.D. from the National Institutes of Health (GM-65982), and the National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, EIA-9802068, and EIA-0305444), and to A.C.A. from the NIH (GM-067542).
1 Introduction

When the structure of an homologous protein is known, initial phases for the diffraction data can often be determined using the technique of molecular replacement (Blow and Rossmann 1962, Crowther and Blow 1967, Rossmann 1990). To use initial phases from an homologous model, each copy of the molecular replacement model (henceforth called model for brevity) must be properly oriented and translated within the asymmetric unit (Tong and Rossmann 1990, Blow and Rossmann 1962, Crowther and Blow 1967, Rossmann 1990). The task of properly orienting and translating each model is facilitated by exploiting the additional constraint provided by non-crystallographic symmetry (NCS).

Traditionally, the initial presence and degree (i.e., three-fold, four-fold, ...) of NCS is first identified by a self-rotation function (Blow and Rossmann 1962), while the orientations of each copy of the model are subsequently identified using the cross-rotation function (Blow and Rossmann 1962). Ideally, after $n$-fold NCS has been identified, the peaks of the cross-rotation function will possess two desirable properties. First, the top $n$ cross-rotation peaks should have rotation-function scores significantly higher than the rest, and second, these top peaks should correspond to the correct NCS-consistent orientations. A complication arises when the model is partial (Oh 1995), when the model shares only moderate structural similarity with the crystallized molecule, or when the degree of NCS is high and each model represents only a small fraction of the unit cell’s molecular mass. Empirically, in these situations, the $n$ top-scoring rotations from a cross-rotation function search may not correspond to the $n$ correct (NCS-consistent) rotations. In fact, some of the correct rotations may not even appear in the cross-rotation peak list, further complicating the search for the correct model orientations. For complex systems it is therefore possible that the wrong cross-rotation peaks are selected for use with the computationally expensive translation function.

We have developed an algorithm to 1) compute which cross-rotation peaks generate NCS-consistent model orientations and 2) generate NCS-consistent model orientations not specified by rotations in the cross-rotation peak list.

The CRANS algorithm identifies and computes sets of rotations (set $\mathcal{R}$ in Figure 1) that, when applied to the model, produce orientations consistent with the NCS. We call these sets of rotations \textit{NCS-consistent rotation sets}. CRANS therefore reduces the time required to obtain initial phases in two ways. First, by correctly identifying NCS-consistent rotation sets among the rotations identified by the cross-rotation function, we reduce the number of improperly-oriented models and thereby the total number of translation searches required to generate initial phases. Second, in the case where one or more NCS-consistent rotations are absent from the cross-rotation peak list, CRANS i) tells the crystallographer that a peak is missing without requiring him/her to perform translation searches on each cross-rotation peak and ii) computes the missing NCS-consistent rotations.

The core of the CRANS algorithm analyzes the set of rotation differences to identify finite sets of rotations that satisfy the subgroup axioms. Given a set of $w$ cross-rotation peaks (rotations) $\mathcal{C} = \{r_1, r_2, \ldots, r_w\}$, where $r_i$ is an element of the group of three-dimensional rotations $SO(3)$, CRANS examines the $w^2$ rotation differences $d_{ij}$ where $d_{ij} = r_i^{-1}r_j$ and $r_i^{-1}$ is the inverse of $r_i$ (such that $r_i^{-1}r_i$ is the identity). Conceptually, the rotation difference $d_{ij}$ is the rotation that rotates the...

---

*Abbreviations used: SO(3), group of three-dimensional rotations; MR, molecular replacement; model, molecular replacement model; DHFR, dihydrofolate reductase; TS, thymidylate synthase; ChDHFR-TS, Cryptosporidium hominis DHFR-TS; LmDHFR-TS, Leishmania major DHFR-TS; PcTS, Pneumocystis carinii TS; NCS, non-crystallographic symmetry; CCP4, collaborative computing project number 4; CNS, crystallography and NMR system; RMSD, root-mean-square distance.
Figure 1: Two dimensional examples for 3-fold (A) and 4-fold (B) NCS. (A) A model is shown with the results of a simplified cross-rotation search $C$ containing only 5 rotations. Orientations corresponding to rotations $r_1$ (purple), $r_3$ (green), and $r_4$ (blue) form an NCS-consistent rotation set $R$. For clarity only a few rotation differences ($d_{11}$, $d_{13}$, and $d_{14}$) are shown in the upper right overlapping orientation figure. The rotation differences $D(R)$ form a complete rotation difference set and satisfy the group properties of associativity (not shown), identity, inverse, and closure. (B) A 4-fold NCS example is shown using similar notation to panel (A). In this example, only 3 of the 4 NCS-consistent rotations ($r_1$, $r_3$, and $r_4$) are contained in the cross-rotation peak list $C$. The missing rotation $r_α = d_{1α}r_1$ is computed using $d_{1α}$ defined by the axis of the three identified NCS-consistent rotations and the missing angle. The now complete NCS-consistent rotation set $R$ has a complete rotation difference set $D(R)$ which satisfies the subgroup properties.

model oriented by $r_j$ into the model oriented by $r_i$ (Figure 1). Therefore, we can test the NCS-consistency of a set of cross-rotation peaks by examining their rotation differences and verifying that they form a finite subgroup of $SO(3)$. In the event of missing rotations, the CRANS algorithm completes each partial set of identified NCS-consistent rotations by generating missing rotations with quaternions. By using quaternions we avoid the well known degeneracies and singularities that occur with most other rotation representations (Drenth 1994, pages 219–220) which could result in incorrect orientations.

When the model $P$ is a homodimer, the search for NCS-consistent rotation sets is more complex because the orientation $rP$ (where $rP$ is the result of rotating protein $P$ by rotation $r$) is equivalent to the orientation $rfP$ where $f$ is the 180° rotation around the dimer 2-fold axis (Figure 2). This rotational degeneracy increases the difficulty of the search because the rotational relationships between the cross-rotation peaks must be examined modulo $f$. Thus the dimer axis is explicitly considered by CRANS when using a homodimer model.

Although the CRANS algorithm only requires a cross-rotation peak list as input, if the NCS axis and/or degree are known (i.e., from the self-rotation search) it can be used as an additional constraint in identifying rotation subgroups. In addition to searching for NCS-consistent rotation sets satisfying one particular symmetry, the CRANS algorithm may also be run in ‘scan’ mode, where the algorithm consecutively attempts to identify subgroups of rotation differences consistent with a user-specified range of NCS degree (i.e., 3-fold, 4-fold, . . . , $n$-fold).

Previous work in automated cross-rotation peak analysis includes the program RFCORR (distributed as part of CCP4 (Collaborative Computational Project Number 4 1994)). Although both
Figure 2: Homodimer-based rotations. The model in orientation A can be rotated by either \( r_1 \) or \( r_2 \) to assume orientations B or C. While both orientations B and C can form a 4-fold NCS (modulo the dimer flip) with orientations X, Y, and Z, the rotations \( d_\alpha \) and \( d_\beta \) are not equal. The rotation difference \( d_\alpha \) is not consistent (i.e., does not form a subgroup) with the partial rotation difference set generated from orientations X, Y, and Z. Thus it becomes important that when the model is a homodimer, for every rotation \( r_1 \) we generate the dimer flip related rotation \( r_2 = r_1 f \) (where the rotation \( f \) flips the model along its 2-fold dimer axis).

RFCORR and CRANS analyze cross-rotation peak lists, CRANS extends the RFCORR algorithm in several ways. RFCORR computes the set of rotation differences but leaves the task of identifying sets of NCS-consistent rotations to the user; CRANS automates this process and identifies NCS-consistent rotation sets even despite missing cross-rotation peaks. When cross-rotation peaks are missing, the CRANS algorithm uses quaternions to compute the missing rotations. The internal rotation representation for the CRANS algorithm is quaternions, while the RFCORR program uses rotation matrices specified by Euler angles. Quaternions have a single compact representation, can be composed by simple multiplication, can easily be normalized to facilitate recovery from accumulated numerical error, and are free from singularities (see Section 2.2). Another novel feature of CRANS is its scan mode: when the NCS-degree is unknown or is known with low-confidence, the CRANS algorithm can quickly scan through a range of degrees of NCS for NCS-consistent rotation sets. Finally, unlike RFCORR, CRANS can identify and generate NCS-consistent rotation sets when the model is a homodimer. In our experience, the ability to handle homodimer models was vital in solving the structure of dihydrofolate reductase-thymidylate synthase (DHFR-TS) from Cryptosporidium hominis (O’Neil et al. 2003) (Section 3.4).

In summary, the CRANS algorithm analyzes the output of a cross-rotation search and makes the following contributions:

1. CRANS computes all sets of \( n \)-fold NCS-consistent rotation sets, even in the presence of potentially missing peaks. The missing peaks (i.e., those not found by the cross-rotation function) required to form a complete (i.e., no missing rotations) NCS-consistent rotation set are also
computed by CRANS. The computed missing peaks can be used to orient the model prior to performing a translation search.

2. During a traditional, manual examination of cross-rotation peaks, one typically only examines the top \( p \) cross-rotation peaks. In contrast, the CRANS algorithm performs an exhaustive search over all peaks in the cross-rotation function, checking each rotation for inclusion into an NCS-consistent rotation set.

3. CRANS computes the NCS axis for each identified NCS-consistent rotation set. If the NCS axis is known from the self-rotation map, CRANS can limit its search to return only those rotation sets with NCS axes close to the known NCS axis.

4. CRANS can also be run in ‘scan’ mode where the algorithm sequentially searches a range of degrees of NCS and attempts to identify NCS-consistent rotation sets for each degree of symmetry.

5. Finally, CRANS can perform the four above listed tasks when the model is a homodimer.

The paper is organized as follows: In Section 2 we describe the CRANS algorithm, explain the benefits of using quaternions, present the algorithm’s runtime complexity, and describe the data and preprocessing used for each test system. In Section 3 the performance of CRANS is demonstrated by analyzing six cross-rotation peak lists, generated from six models against four protein crystal systems. First, we describe the performance of the CRANS algorithm on three test systems of solved crystal structures, containing 3-fold (2-Keto-3-Deoxy-6-Phosphogluconate Aldolase (1FQ0) (Wymer et al. 2001)), 5-fold (Cholera Toxin B Subunit (1CHP) (Merritt et al. 1995)), and 7-fold (Gp31 Co-Chaperonin (1G31) (Hunt et al. 1997)) NCS. Next, in Section 3.4 we describe how we used CRANS to solve the \textit{de novo} structure of dihydrofolate reductase-thymidylate synthase (DHFR-TS) from \textit{Cryptosporidium hominis} (ChDHFR-TS) (1QZF) (O’Neil et al. 2003), a homodimer, where we used CRANS to identify the cross-rotation peaks consistent with 5-fold NCS. Despite the fact that three of the six tested cross-rotation peak lists were missing at least one and in one instance as many as three NCS-consistent rotations, in all cases, CRANS was able to correctly identify NCS-consistent rotations and generate all missing rotations. These results support the general applicability of the CRANS algorithm for use in analyzing cross-rotation search results for systems with NCS.

2 Methods

2.1 Subgroup Search

We first formalize the search for NCS-consistent rotation sets as a search for finite subgroups of \( SO(3) \), showing that the search for finite subgroups of \( SO(3) \) among the rotation differences of cross-rotation peaks provably identifies NCS-consistent rotation sets. We then illustrate the algorithm with an example.

The \textit{rotation difference set} of a set of rotations \( R \) is defined as:

\[
D(R) = \{r_i^{-1}r_j \mid r_i, r_j \in R\}
\] (1)
(where equal rotations are identified). Recall that \( r_j^{-1}r_i \) is the rotation difference \( d_{ij} \) between rotations \( r_i \) and \( r_j \). When \( D(R) \) is a subgroup of \( SO(3) \), we say that \( R \) and \( D(R) \) are complete. When \( R \) is complete, a partial rotation difference set is a subset \( D(R') \subset D(R) \) constructed from a subset of rotations \( R' \) of \( R \). Therefore a partial rotation difference set can be made complete by adding missing rotation differences to complete the subgroup.

The goal of our algorithm is: given a set of cross-rotation peaks \( C \subset SO(3) \), find finite subsets of rotations \( R \subset C \) such that \( D(R) \) is a subgroup of \( SO(3) \).

We now show that an NCS-consistent rotation set \( R \) will have a complete rotation difference set \( D(R) \) satisfying the subgroup axioms, and, conversely that a rotation set \( R \) with a complete rotation difference set \( D(R) \) that satisfies the subgroup axioms, is an NCS-consistent cross-rotation peak set.

For \( D(R) \) to be a subgroup the four group properties of identity, inverse, associativity, and closure must be satisfied. All complete rotation difference sets satisfy the properties of identity, inverse, and associativity. By definition (Eq. 1) the rotation difference \( d_{ii} = I \) (where \( I \) is the identity element) is present in all complete rotation difference sets. The inverse property for each rotation difference \( d_{ij} \) from \( r_i \) to \( r_j \) is satisfied by the rotation from \( r_j \) to \( r_i \), that is, \( d_{ij}d_{ji} = I \). Finally, all rotations compose associatively by construction. Thus the crux of the proof is in the closure property. To prove the forward direction, we note that all rotation differences are a multiple of \( 360/n \) degrees around a common axis (where \( n \) is the degree of NCS) and that all \( n \) rotations \( 360t/n \) \( (t \in \{0, 1, \ldots, n-1\}) \) are represented in the set \( D(R) \). Thus the application of any number of rotation differences results in a rotation that (modulo \( 360^\circ \)) is already in the complete rotation difference set. Conversely, to prove the reverse direction, all finite subgroups have an identity and satisfy the closure property (modulo \( 360^\circ \)); therefore, if they contain \( n \) members they represent \( n \)-fold symmetry and the symmetry angles \( 360t/n \). This completes the proof.

When the model is a homodimer, each cross-rotation peak and its corresponding dimer-related symmetry mate produce equivalently-oriented dimers. Therefore each cross-rotation peak and its symmetry mate must be used to form finite subgroups and subsequently NCS-consistent rotation sets.

We now present a synthetic two-dimensional 3-fold NCS example to illustrate the framework presented in this section. Figure 1A shows the model and five resulting cross-rotation function peaks \( C = \{r_1, r_2, r_3, r_4, r_5\} \). The set of peaks \( R = \{r_1, r_3, r_4\} \) forms an NCS-consistent rotation set. The complete rotation difference set \( D(R) \) contains three rotations around an NCS-axis coming out of the page, \( d_\alpha \) corresponds to the identity rotation, \( d_\beta \) corresponds to a 120-degree clockwise or 240-degree counter-clockwise rotation, and \( d_\gamma \) corresponds to a 240-degree clockwise or a 120-degree counter-clockwise rotation. The subgroup \( D(R) \) satisfies the four group properties (see Figure 1A): the group property of associativity is satisfied by all rotations, \( d_\alpha \) is the identity rotation and also serves as its own inverse while \( d_\beta \) is the inverse of \( d_\gamma \) and \( d_\gamma \) is the inverse of \( d_\beta \). Finally, the closure property is satisfied, \( d_\alpha d_\beta = d_\beta d_\alpha = d_\beta \), \( d_\alpha d_\gamma = d_\gamma d_\alpha = d_\gamma \), and \( d_\beta d_\gamma = d_\gamma d_\beta = d_\alpha \).

2.2 Quaternions

When performing rotation operations, our algorithm uses quaternions (Hamilton 1969, Salamin 1979). Quaternions have a single accepted definition with a compact representation and a number of desirable properties. First, the single quaternion definition simplifies the confusion introduced by, for example, the over twenty ‘standard’ Euler angle representations. Second, quaternions can be composed by simple multiplication: the quaternion \( q_3 = q_2q_1 \) represents the sequential application
of rotation \( q_1 \) followed by \( q_2 \). Third, while accumulation of numerical errors will cause all rotation representations to deviate from pure rotations, quaternions have the advantage that they can efficiently be returned to a pure rotation by simply normalizing the vector. Finally, quaternions represent a uniform parameterization of rotation space, thereby avoiding the problems encountered with non-uniform parameterizations which contain singularities (i.e., Euler angles and axis-angle representations) (Drenth 1994, pages 219–220). Small deviations in the parameters of non-uniform rotation representations can lead to very different rotations. All these properties make quaternions the representation of choice for working with and composing rotations in our algorithm.

2.3 CRANS Algorithm

The CRANS algorithm takes a list of cross-rotation peaks as input and identifies sets of cross-rotation peaks that generate orientations of the model related by \( n \)-fold NCS.

In describing the CRANS algorithm we define the functions \( \text{axis}(\cdot) \) and \( \text{angle}(\cdot) \) to return the axis and angle, respectively, of a given rotation. We define the \( n \) angles specified by \( 360t/n \) \((t \in 0, 1, \ldots, n - 1)\) as symmetry angles. Let \( m \) be the maximum number of missing rotations allowed by the user.

The CRANS algorithm is divided into four stages: process input, filter, partition, and patch. In the process input step, all \( w \) original rotations (rotations from the cross-rotation function) \( C = \{r_1, r_2, \ldots, r_w\} \) are read and converted to quaternions. Then, all rotation differences \( d_{ij} \) are computed.

In the filter stage, all rotation differences with \( \text{angle}(d_{ij}) \) differing from a symmetry angle by more than the user-defined angular threshold \( \tau_{\text{angle}} \) are discarded. The intuition behind this filtering step is that two rotations \( r_i \) and \( r_j \) that are both members of the correct NCS-consistent rotation set should have a rotation difference \( \text{angle}(d_{ij}) \) approximately equal to a symmetry angle.

The core of the algorithm occurs in the partition stage. In general, if rotation differences \( d_{ij} \) and \( d_{kl} \) are both members of the same NCS-consistent rotation set, then \( \text{axis}(d_{ij}) \approx \text{axis}(d_{kl}) \). Therefore, in this stage, each remaining (i.e., unfiltered by step 2) rotation difference \( d_{ij} \) is clustered with all rotation differences \( d_{ik} \) that share a common rotation axis. That is, the differences \( d_{ij} \) and \( d_{ik} \) are assigned to the same set if \( 1 - |\cos \theta_{ij,ik}| \leq \tau_{\text{axis}} \), where \( \theta_{ij,ik} \) is the angle between \( \text{axis}(d_{ij}) \) and \( \text{axis}(d_{ik}) \) and \( \tau_{\text{axis}} \) is the user-defined axis similarity threshold. Sets of rotation differences with more than \( m \) missing distinct rotations are eliminated. In every set \( S_i \) of rotation differences of the form \( d_{ij} \) (where \( i \) is fixed and \( j \) varies, i.e., all rotation differences involve rotation \( r_i \)) the number of distinct rotations equals the number of unique orientations generated by applying \( d_{ij} \in S_i \) to \( r_iP \) where \( P \) is the model. Because multiple rotations in a single rotation difference set might generate the same orientation, all possible subsets of rotation differences containing at most \( m \) missing distinct rotations are computed. At this point, each set \( S_i \) of rotation differences has generated zero or more partial rotation difference sets \( \Delta' \). For each partial rotation difference set \( \Delta' \), let \( D^{-1}(\Delta') \) be the set of all original cross-rotation peaks used to generate rotations in \( \Delta' \). The consistency of the NCS-axis for each rotation difference is checked by comparing the two rotation differences \( d_{ab} \) and \( d_{bc} \) for all triples \( r_a, r_b, r_c \in D^{-1}(\Delta') \) to ensure \( 1 - |\cos \theta_{ab,bc}| \leq \tau_{\text{axis}} \) where as before, \( \theta_{ab,bc} \) is the angle between \( \text{axis}(d_{ab}) \) and \( \text{axis}(d_{bc}) \). Rotation difference sets not passing this filter are eliminated. After computing all rotation sets, those sets that are subsets of other, more complete, remaining rotation sets are removed.

In the final stage of the algorithm, the patch stage, missing rotations for each remaining rotation set \( R \) are computed. For example (Figure [B]), if \( n = 4, m = 1 \), and a rotation set \( R \) contains
three rotations $r_1$, $r_3$, and $r_4$ such that \( \text{axis}(d_{13}) = \text{axis}(d_{14}) = \text{axis}(d_{34}) \) and \( \text{angle}(d_{13}) = 90^\circ, \text{angle}(d_{14}) = 270^\circ, \text{angle}(d_{34}) = 180^\circ \), then the missing rotation $r_\alpha = d_{1\alpha}r_1$ is computed where $d_{1\alpha}$ is the quaternion representing a rotation of $180^\circ$ (the missing angle in this example) around $\text{axis}(d_{13})$ (Figure 1B). By construction, $D(R)$ will be a subgroup and thus $R$ will be an NCS-consistent rotation set.

When the model is a homodimer, one change is made to the process input stage of the above algorithm. First, the dimer axis is computed from the model structure and is used to compute $f$, the $180^\circ$ rotation around the 2-fold dimer axis (Figure 2). The rotation $f$ is then used to generate a dimer-flipped version of each original rotation which is added to the list of rotations utilized by subsequent stages of the algorithm. Formally, the set of cross-rotation peaks $C$ is replaced by $C \cup \{rf \mid r \in C\}$. The algorithm then proceeds as described above.

2.4 Complexity

Reading, conversion to quaternions, and computation of the inverse of each of the $w$ cross-rotation peaks requires $O(w)$ time. Computing all rotation pair differences requires $O(w^2)$ time. Filtering the rotation pair differences (for those near the $n$ symmetry angles) requires time $O(nw^2)$. Formation of rotation difference sets sharing a common axis takes a worst case time of $O(w^4)$, which occurs when all $w^2$ rotation differences have passed the previous symmetry angle filter. In practice, the number of remaining differences is quite small and only a fraction of the original rotation differences are used to compute rotation difference sets. Creation and patching of the final sets consistent with the desired NCS, requires time $O(g)$ where $g$ is the number of NCS-consistent subgroups. Therefore, the CRANS algorithm can search for NCS symmetry among $w$ cross-rotation peaks with an expected runtime of $O(nw^2 + g)$ and a worst case runtime of $O(w^4 + g)$. In practice, $g$ is a small constant and we can reduce the runtime to expected $O(nw^2)$ and worst case $O(w^4)$. Our implementation of CRANS requires only seconds to search for up to 8-fold NCS on lists of 120 cross-rotation peaks using an Athlon-based processor.

2.5 Source of Data and Preprocessing

Structure factors for the three test systems 1FQ0, 1CHP, and 1G31 were obtained from the protein databank (PDB) (Berman et al. 2000) and converted into CNS (Brunger et al. 1998) format. The first chain from the crystal structure (representing a single monomer) was extracted and used as the model for molecular replacement. A cross-rotation search was performed with CNS using default parameters: Resolution limits: 15-4 Å, Data cutoff criteria: 0.0, RMS outlier cutoff: 1000, Bins for resolution-dependent operations: 10, Atoms: known and not hydrogen, Scoring function: fastdirect, Use automatically determined asymmetric unit, Fastdirect grid factor: 5, Fastdirect maximum number of peaks: 20, Cluster threshold: 10°, Max number of coarse peaks to analyze in fine grid search: 20. The generated cross-rotation function peak list was then analyzed by CRANS.

The same axis and angle tolerances of $\tau_{\text{axis}} = 0.003$ and $\tau_{\text{angle}} = 5.0^\circ$ were used in all CRANS analyses.

For comparison, the translation-only RMSD (TRMSD) was computed between the model oriented according to each cross-rotation peak and each chain in the crystal structure. We define the TRMSD of two proteins to be the minimum main chain C$_{\alpha}$ RMSD achievable when the molecules are only allowed to translate relative to one another (i.e., rotations are not allowed). Using this measure we are able to quantify the similarity between models oriented by each cross-rotation peak
and each monomer of the crystal structure. We emphasize that the TRMSD is not part of the CRANS algorithm but rather is a tool used to analyze the CRANS output to verify the correctness of our test cases.

Diffraction data for ChDHFR-TS were collected at Brookhaven National Lab (beamline X12C) as previously described (O’Neil et al. 2003). Experiments were performed using three molecular replacement models. The first model, LmDHFR-TS, consists of the DHFR-TS homodimer of the *Leishmania major* DHFR-TS protein (Knighton et al. 1994). The second model, PcTSA, consists of the TS homodimer of the *Pneumocystis carinii* TS protein (PDB: 1F28) (Anderson et al. 2001) while the third model, PcTSB, is simply a rotated version of PcTSA with the flexible loop Asn186-Glu191 removed. Both the PcTSA and PcTSB models only consist of the TS homodimer which represents approximately 60% of the entire DHFR-TS homodimer.

### 3 Results

The CRANS algorithm was tested on structure factor data from four different protein crystals exhibiting 3-, 5-, and 7-fold NCS. Three of these systems exhibit planar NCS whereas ChDHFR-TS exhibits a $5_1$ screw NCS (Figure 3).

#### 3.1 2-Keto-3-Deoxy-6-Phosphogluconate Aldolase (1FQ0) (3-fold NCS)

Structure factors for 2-keto-3-deoxy-6-phosphogluconate aldolase (1FQ0) were obtained from the PDB. The first chain of 1FQ0 was used as a model in a cross-rotation search using default parameters (Section 2.5). The resulting cross-rotation peak list was sorted by rotation-function score and
contained 162 rotations. The TRMSD (Section 2.5) was computed between the model oriented by each cross-rotation peak and each chain of the 1FQ0 crystal structure. While the average and standard deviation (in parentheses) TRMSD between each cross-rotation peak oriented model and the closest chain of the refined structure is 16.4 (7.2) the closest (i.e., smallest TRMSD) peak to chain A has a TRMSD of 0.01 (peak 1), to chain B has a TRMSD of 4.00 (peak 137), and to chain C has a TRMSD of 3.44 (peak 46). In this test case, although peak 1 corresponds to the orientation closest to chain A, the top three rotation-function ranked cross-rotation peaks do not correspond to correct model orientations. The crans algorithm was directed to search for complete 3-fold NCS (no missing peaks) and identified peaks 1 (chain A TRMSD 0.01), 159 (chain B TRMSD 6.28) and 59 (chain C TRMSD 4.00) as the highest scoring NCS-consistent rotation set (Table 1). While the NCS-consistent rotation set consisting of peaks 1, 137, and 46 contains those cross-rotation peaks producing orientations with the smallest TRMSD to the refined structure, these orientations do not strictly obey the NCS as tightly as peaks 1, 159, and 59 (originally found by crans). The orientations specified by cross-rotation peaks 1 and 137 have a relative angle of 108.6° rather than the NCS specified 120°. Therefore, as a control, we relaxed the values of \(\tau_{\text{axis}}\) and \(\tau_{\text{angle}}\) and reran crans. With the relaxed thresholds, crans was able to find the NCS-consistent rotation set consisting of peaks 1, 137, and 46. Despite the fact that the overall quality of the rotations returned by the 1FQ0 cross-rotation function are low compared to the cross-rotation peaks found in the other test cases (1CHP, 1G31, and ChDHFR-TS), crans was still able to extract 3 rotations that result in properly oriented models.

The column ‘Best Pk. in top 10’ of Table 1 lists the cross-rotation peak with the smallest (best) TRMSD among the top 10 rotation-function ranked cross-rotation peaks. This column lists the peak which might be found in a manual molecular replacement effort. Because crans exhaustively checks all peaks in the cross-rotation list, NCS-consistent rotation peaks that appear at the bottom of the cross-rotation list are found as easily as those that appear near the top of the list.

### 3.2 Cholera Toxin B Subunit Mutant (1CHP) (5-fold NCS)

Structure factors for Cholera Toxin B Subunit Mutant (1CHP) were obtained from the PDB. The first chain of 1CHP was used as a model in a cross-rotation search using default parameters. The resulting cross-rotation peak list was sorted by rotation-function score and contained 122 rotations. The TRMSD was computed between the model oriented by each cross-rotation peak and each chain of the 1CHP crystal structure. In this case, unlike 1FQ0, the top 5 cross-rotation peaks did have the lowest TRMSD to each of the 5 chains in the crystal structure, respectively 0.10, 0.62, 1.02, 0.68 and 0.54. The crans algorithm was directed to search for complete 5-fold NCS (no missing peaks) and identified two sets of cross-rotation peaks. The first set contained peaks 1, 2, 3, 4, and 5 which correspond to the peaks with the lowest TRMSD to the refined structure (Table 1). crans also identified a second complete set of cross-rotation peaks containing peaks 59 (chain D TRMSD 1.96), 75 (chain E TRMSD 2.83), 58 (chain F TRMSD 2.83), 22 (chain G TRMSD 1.91), and 45 (chain H TRMSD 1.97). Thus both identified rotation sets are consistent with the crystal structure.

### 3.3 Gp31 Co-Chaperonin from Bacteriophage T4 (1G31) (7-fold NCS)

Structure factors for Gp31 co-chaperonin from bacteriophage T4 (1G31) were obtained from the PDB. The first chain of 1G31 was used as a model in a cross-rotation search using default param-
Table 1: TRMSDs measured for the three test systems, 1FQ0, 1CHP, and 1G31. Column 2 lists the PDB chain identifier used in computing the TRMSD for the specified row. The ‘Best Pk.’ column lists the minimum TRMSD observed between the specified chain and all cross-rotation peaks along with its corresponding cross-rotation peak index (in parentheses). Cross-rotation peaks are ordered by sorting them based on the cross-rotation function score where a lower index corresponds to a higher (better) cross-rotation score. The TRMSD and peak index of the peak with the smallest TRMSD among the top 10 cross-rotation function ranked peaks is listed in the ‘Best Pk. in top 10’ column. The TRMSD of the crans identified peak of the NCS-consistent rotation set corresponding to the specified chain and its cross-rotation peak index (in parentheses) is listed in column ‘crans Pk.’. Peaks computed by crans (i.e., those missing in the cross-rotation peak list) are shown in italics with a dash for the peak index.
3.4 Cryptosporidium hominis DHFR-TS (5-fold NCS)

Molecular replacement was used to determine initial phase angles for the structure of ChDHFR-TS. Diffraction data to 2.8 Å were collected at Brookhaven National Laboratory (beamline X12C) and processed into structure factors as previously described (O’Neil et al. 2003). Analysis of self-rotation peaks indicated the presence of 5-fold NCS. A cross-rotation search using default parameters was performed with each of three models (LmDHFR-TS, PcTSA, and PcTSB; see Section 2.5). The resulting cross-rotation peak lists were sorted by rotation-function score and run through the crans algorithm (because each model was a homodimer, crans was run in homodimer-mode). The LmDHFR-TS, PcTSA, and PcTSB cross-rotation peak lists contained 45, 38, and 38 rotations respectively. Cross-rotation peak analysis with crans was able to find 5-fold NCS sets with 1 missing peak for LmDHFR-TS, 0 missing peaks for PcTSA, and 2 missing peaks for PcTSB. The NCS axes computed for all crans identified NCS-consistent rotation sets agreed with the axis identified by the self-rotation search. Models oriented according to the rotations of the complete PcTSA NCS-consistent rotation set were positioned using a translation search (O’Neil et al. 2003, Crowther and Blow 1967). The initial R-factor of 52% was refined to 22.5% ($R_{\text{free}} = 24.5\%$). Refined ChDHFR-TS molecules have a non-crystallographic 5$_1$ axis (O’Neil et al. 2003), see Figure 3.

Analysis performed after the structure determination clearly explains the crans results. The TRMSD was computed between the thymidylate-synthase (TS) domain of each model oriented by each cross-rotation peak and each TS homodimer (dimers A, B, C, D, and E) of the refined ChDHFR-TS crystal structure. Although the DHFR domains were used in the LmDHFR-TS cross-rotation search, the DHFR domains were not used in computing the TRMSDs because of the significant difference in the refined DHFR orientations relative to the highly-conserved TS homodimer. The TRMSDs are presented in Table 2.

**LmDHFR-TS.** TRMSD analysis of the cross-rotation search peaks show that while some peaks closely approximate the structures of homodimers A, B, C, and D, no cross-rotation peaks returned were similar to homodimer E (Table 2). Despite the significant structural and relative orientational differences between the DHFR domains of LmDHFR-TS and ChDHFR-TS, four of five correct rotations were still found in the LmDHFR-TS cross-rotation search using default search parameters. Table 2 shows that peaks 2 and 5 provide redundant information and both correspond to homodimer C (and its dimer flip). Therefore, the naïve selection of the top 5 cross-rotation peaks, in the hope that they corresponded to the 5 correct NCS related rotations, would result in only four of five correct orientations and one redundant orientation. While peaks 2 and 5 corresponded to homodimer C and its dimer flip, the only cross-rotation peaks corresponding to homodimers A, B, and D matched either the dimer or its dimer flip (but not both). Consequently, the ability of the crans algorithm to handle homodimer models was vital in this analysis. Specifically, a search for rotation difference subgroups that ignored the fact that the model is a homodimer would not have been able to identify all 4 NCS related peaks contained in the cross-rotation peak list. The peaks identified by crans correspond to those peaks with the smallest TRMSD for each homodimer (Table 2). Furthermore, the rotation generated by crans to complete the NCS-consistent rotation set has a TRMSD to homodimer E of 1.33 Å.

**PcTSA.** Unlike the LmDHFR-TS model, PcTSA consists of only a TS homodimer. Cross-rotation peak analysis with crans found one 5-fold NCS-consistent rotation set with no missing peaks (Table 2). The peaks identified by crans correspond to those with the smallest TRMSD to each of the ChDHFR-TS homodimers. As with LmDHFR-TS, the ability to handle homodimer
Table 2: TRMSDs measured for the three homologous models used in solving the ChDHFR-TS structure. Column 2 lists the identifier of the homodimer used in computing the TRMSD for the specified row. The ‘Closest Pk.’ and ‘crans Pk.’ columns are as listed in the caption of Table 1. The ‘Closest Dimer Flipped Pk.’ column is similar to the ‘Closest Pk.’ column however TRMSDs are computed between the dimer flip of the specified homodimer and each peak in the cross-rotation peak list. Because crans computes model rotations invariant to homodimer flips, the crans computed TRMSDs are taken as the smaller of the TRMSD to the crystallographic dimer or its dimer flip.

models was crucial in crans analysis.

PcTSB. Similar to the PcTSA model, PcTSB consists of only a TS homodimer (see page 8). Analysis of the cross-rotation peak sets with crans could not identify a 5-fold NCS-consistent rotation sets with zero or one missing peaks, however four NCS-consistent rotation sets were found with peaks corresponding to 3 of the 5 TS homodimers. Because peaks 1 and 2 (resp. 3 and 4) provide redundant information and both correspond to orientations of homodimer B (resp. homodimer C) of the refined structure, the four identified sets correspond to the four sets \{1, 3, 5\}, \{2, 3, 5\}, \{1, 4, 5\}, \{2, 4, 5\} \equiv \{1, 2\} \times \{3, 4\} \times \{5\} (numbers are cross-rotation peak numbers sorted by rotation-function score). The top scoring set was \{1, 3, 5\} which has TRMSDs of 1.21Å, 1.15Å, 1.19Å to the crystallographic ChDHFR-TS homodimers. The two computed (missing peaks) for this set have TRMSDs of 1.32Å and 1.34Å (Table 2). Although the top five cross-rotation peaks had a rotation-function score approximately twice that of the remaining peaks, the direct application of the top five peaks would not have resulted in the five orientations seen in the refined structure. Therefore crans analysis provided important information that not all five of the orientations seen in the refined structure were seen among the top five peaks of the cross-rotation list.

In summary, although two of the three cross-rotation peak lists did not contain all 5 NCS-consistent model rotations, the crans algorithm was successfully able to 1) verify that 5-fold NCS was present, 2) find sets of cross-rotation peaks related by an NCS-axis consistent with the self-rotation function for all three models, and 3) compute missing cross-rotation peaks corresponding to orientations with TRMSDs of 1.33Å, 1.34Å, and 1.32Å to the final crystal structure. The identified NCS-consistent rotation sets were then used in a translation function and rigid body refinement to generate initial phase angles and the final structure of ChDHFR-TS (1QZF) (O’Neil et al. 2003).
The CRANS-identified NCS-consistent rotation sets were correct and agree with the final refined structure.

4 Discussion

In all six cross-rotation function searches performed, only the default search parameters were used. That is, we did not spend any time optimizing cross-rotation search parameters. It is possible that by tweaking cross-rotation function search parameters and by optimizing the model (i.e., removing flexible loops, changing residues to Ala, etc...) that more NCS-consistent rotations could have been returned by the cross-rotation function. The time required to perform these optimizations can be reduced or eliminated by using CRANS to analyze the output of a cross-rotation search run with default parameters.

As the number of peaks returned by the cross-rotation function increases, the probability that these peaks will conspire to form ‘fake’ low-degree NCS-consistent rotation sets rises. This is especially true if the axis and angle tolerances are not particularly tight. Therefore, when analyzing low-degree NCS with large cross-rotation peak lists, the results of a CRANS search should be treated as a working hypothesis. Confidence in CRANS-identified NCS-consistent rotation sets can be increased by directing CRANS to use the NCS axis identified by the self-rotation map.

CRANS identifies NCS-consistent rotation sets from a cross-rotation search using monomer or homodimer models, since these are the most common model types used in molecular replacement; however, the CRANS algorithm can be extended to handle any degree of oligomerization (e.g., models that are homotrimers, homotetramers, or homopentamers). Conceptually, to handle higher order model symmetry, all symmetry rotations of the model are computed and then applied to each cross-rotation peak. For example, with $d$-fold model symmetry ($d$-fold oligomerization), let $f$ be the $360/d$-degree rotation around the symmetry-axis. Each rotation $r$ identified by the cross-rotation function is replaced by the $d$ rotations $rf^z$ ($z \in 0, 1, \ldots, d - 1$). Heterodimers do not inherently contain a symmetry axis and therefore avoid the rotational degeneracy presented by homodimers. Thus any model which does not have a symmetry axis (e.g., heterodimers, heterotrimers) should be treated as a monomer by CRANS.

While the tests in this paper were performed with an ordinary cross-rotation function, an alternative is to use a locked cross-rotation function (Tong and Rossmann 1990, Tong 2001) when the NCS axis is clear from the self-rotation function. Although the scoring of the rotation function peaks is different, in the presence of NCS, both the ordinary and locked cross-rotation functions should identify rotations corresponding to each of the $n$ NCS-consistent orientations. When using a locked rotation function, non-crystallographic symmetry mates are generated using the NCS axis initially identified by the self-rotation function. However, if the NCS axis cannot be defined with a high degree of accuracy or the conservation of non-crystallographic symmetry in the crystal is not perfect, then a more accurate orientation for each monomer may be identified by orienting each monomer individually. CRANS can process the results of ordinary or locked cross-rotation functions to return a set of NCS-consistent orientations. The NCS-consistent model orientations identified and generated by CRANS can be positioned either sequentially using an ordinary translation function (Crowther and Blow 1967) or simultaneously using a locked translation function (Tong and Rossmann 1990, Tong 2001) followed by an ordinary translation function (Tong 2001).
5 Conclusions

In this paper we presented the CRANS algorithm for analyzing lists of cross-rotation peaks both to extract NCS-consistent rotation sets and to complete partial sets by computing missing NCS-consistent rotations. We showed that the problem of identifying NCS-consistent rotation sets is equivalent to subgroup identification among differences in rotations of the cross-rotation peak list. We then tested the algorithm on four test proteins displaying 3-, 5-, and 7-fold NCS using six models for molecular replacement (three from the solved crystal structure itself and three from homologous proteins). We demonstrated the ability of the CRANS algorithm to find NCS-consistent rotation sets both when all appropriate rotations were present and when up to three rotations (in the case of 1G31) were missing. Furthermore, CRANS successfully identified orientations that were used to generate initial phases in solving the structure of ChDHFR-TS. For all test cases, the CRANS algorithm was able to successfully generate correct NCS-consistent rotation sets. The CRANS algorithm is efficient, requiring only seconds on an Athlon-based processor to search for up to 8-fold NCS on lists of 120 cross-rotation peaks.

By extracting more information from each cross-rotation peak list, the CRANS algorithm provides two main benefits to the crystallographer. First, in the case where the cross-rotation peak list contains \( n \) peaks with significantly higher scores than the remainder of the list (where \( n \) is the NCS-degree), the CRANS algorithm can confirm that these top peaks are indeed consistent with known NCS. The importance of this confirmation was demonstrated in the 1FQ0, 1G31, and ChDHFR-TS systems where the top \( n \) peaks did not correspond to the correct NCS-consistent orientations, despite a frequently sharp dropoff in rotation-function score after these top peaks. Second, when the model is partial or the NCS-degree is high, it becomes likely that one or more NCS-consistent rotations will not be present in the cross-rotation peak list. When this happens, it is not possible to find a complete list of NCS-consistent cross-rotation peaks. The CRANS algorithm can find partial NCS-consistent rotation sets and then generate the missing NCS-consistent rotations to create a complete set. The ability to correctly generate missing peaks was demonstrated in the 1G31, LmDHFR-TS, and PcTSB cases. By using quaternions to generate missing rotations we avoid rotational instability (i.e., singularities) which can arise when using other rotation representations.

6 Supporting Materials

The CRANS program is distributed as a java jar file and is available at http://www.cs.dartmouth.edu/~brd/Bio and by contacting the authors. The software is distributed under the Gnu Public License (Gnu 2002).

7 Acknowledgments

We thank Ramgopal Mettu, Robert O’Neil, Anthony Yan, Robert Stroud, and all members of the Donald and Anderson Labs for helpful discussions and comments on drafts.

References


