

DEPARTMENT OF COMPUTER SCIENCE
SERIES OF PUBLICATIONS C
REPORT C-2004-3



**The Computational Complexity
of Orientation Search Problems
in Cryo-Electron Microscopy**



Taneli Mielikäinen Janne Ravantti

Esko Ukkonen



UNIVERSITY OF HELSINKI
FINLAND

The Computational Complexity of Orientation Search Problems in Cryo-Electron Microscopy

Taneli Mielikäinen

Department of Computer Science
University of Helsinki, Finland
tmielika@cs.Helsinki.FI

Janne Ravantti

Institute of Biotechnology and
Faculty of Biosciences
University of Helsinki, Finland
ravantti@cs.Helsinki.FI

Esko Ukkonen

Department of Computer Science
University of Helsinki, Finland
ukkonen@cs.Helsinki.FI

Department of Computer Science, University of Helsinki
Technical report, Series of Publications C, Report C-2004-3
Helsinki, June 2004, ii + 21 pages

Abstract

In this report we study the problem of determining three-dimensional orientations for noisy projections of randomly oriented identical particles. The problem is of central importance in the tomographic reconstruction of the density map of macromolecular complexes from electron microscope images and it has been studied intensively for more than 30 years.

We analyze the computational complexity of the orientation problem and show that while several variants of the problem are *NP*-hard, inapproximable and fixed-parameter intractable, some restrictions are polynomial-time approximable within a constant factor or even solvable in logarithmic space. The orientation search problem is formalized as a constrained line arrangement problem that is of independent interest. The negative complexity results give a partial justification for the heuristic methods used in orientation search, and the positive complexity results on the orientation search have some positive implications also to the problem of finding functionally analogous genes.

A preliminary version “The Computational Complexity of Orientation Search in Cryo-Electron Microscopy” appeared in Proc. ICCS 2004, LNCS 3036, pp. 231–238. Springer-Verlag 2004.

Computing Reviews (1998) Categories and Subject Descriptors:

- F.2.2 Analysis of Algorithms and Problem Complexity: Nonnumerical Algorithms and Problems
- I.4.5 Image Processing and Computer Vision: Reconstruction
- J.3 Life and Medical Sciences: Biology and Genetics

General Terms:

Algorithms, Theory

Additional Key Words and Phrases:

Orientation Search, Line Arrangement, *NP*-hard, Inapproximable, Fixed-Parameter Intractable, Cryo-Electron Microscopy, Structural Biology

1 Introduction

Structural biology studies how biological systems are built. Especially, determining three-dimensional electron density maps of macromolecular complexes, such as proteins or viruses, is one of the most important tasks in structural biology [15].

Standard techniques to obtain three-dimensional density maps of such particles (at atomic resolution) are by X-ray diffraction (crystallography) and nuclear magnetic resonance (NMR) studies. However, X-ray diffraction requires that the particles can form three-dimensional crystals and the applicability of NMR is limited to relatively small particles [8]. For example, there are many well-known viruses that do not seem to crystallize and are too large for NMR techniques. (To the best of our knowledge NMR techniques can be currently applied only up to size of 1 MDa [14] while viruses are typically at least ten times larger.)

A more flexible way to reconstruct density maps is offered by cryo-electron microscopy [10, 15]. Currently the resolution of the cryo-electron microscopy reconstruction is not quite as high as resolutions obtainable by crystallography or NMR but it is improving steadily.

Reconstruction of density maps by cryo-electron microscopy consists of the following subtasks:

Specimen preparation. A thin layer of water containing a large number of identical particles of interest is rapidly plunged into liquid ethane to freeze the specimen very quickly. Quick cooling prevents water from forming regular structures [15]. Moreover, the particles get frozen in random orientations in the iced specimen.

Electron microscopy. The electron microscope produces an image representing a two-dimensional projection of the iced specimen. This image is called a *micrograph*. Unfortunately the electron beam of the microscope rapidly destroys the specimen so getting accurate images from it is not possible.

Particle picking. Individual projections of particles are extracted from the micrograph. There are efficient methods to do that, see e.g. [21, 25]. The number of projections obtained may be thousands or even more.

Orientation search. The orientations (i.e., the projection directions for each extracted particle) for the projections are determined. There are a few heuristic approaches for finding the orientations. For further details, see Section 2.

Reconstruction. If the orientations for the projections are known then quite

standard tomography techniques can be applied to construct the three-dimensional electron density map from the projections [15].

For a more broader view to the reconstruction process, see Figure 1.

In this report we study the computational complexity of the orientation search problem which is currently the major bottleneck in the reconstruction process. On one hand we show that several variants of the task are computationally very difficult. This justifies (to some extent) the heuristic approaches used in practice. On the other hand we give exact and approximate polynomial-time algorithms for some special cases of the task that are applicable e.g. to the seemingly different task of finding functionally analogous genes [17].

The rest of this report is organized as follows. In Section 2 the orientation search problem is described. Section 3 analyzes the computational complexity and approximability of the orientation search problem. As an abstract formulation of the search problem we use certain constrained line arrangement problems that are of independent interest. The report is concluded in Section 4.

2 The Orientation Search Problem

A *density map* is a mapping $D : \mathbb{R}^3 \rightarrow \mathbb{R}$ with a compact support. An *orientation* o is a rotation of the three-dimensional space and it can be described e.g. by a three-dimensional rotation matrix.

A *projection* p of a three-dimensional density map D to orientation o is the integral

$$p(x, y) = \int_{-\infty}^{\infty} D \left(R_o [x, y, z]^T \right) dz$$

where R_o is a three-dimensional rotation matrix, i.e., the mass of D is projected on a plane passing through the origin and determined by the orientation o .

Projections of physical densities can be produced e.g. by X-rays or electron microscopy. In practice, the density maps are usually represented as three-dimensional regular grids of finite-precision numbers depending on the accuracy of the scanning device but in this report we do not need to consider the actual representations of projections or density maps.

Based on the above definitions, the orientation search task is, given projections p_1, \dots, p_n of the same underlying but unknown density map D to find good orientations o_1, \dots, o_n for them. There are several heuristic definitions of what are the good orientations for the projections.

One possibility is to choose those orientations that determine a good density map although it might not be obvious what a good density map is nor how it should be constructed from oriented projections. A standard solution is to

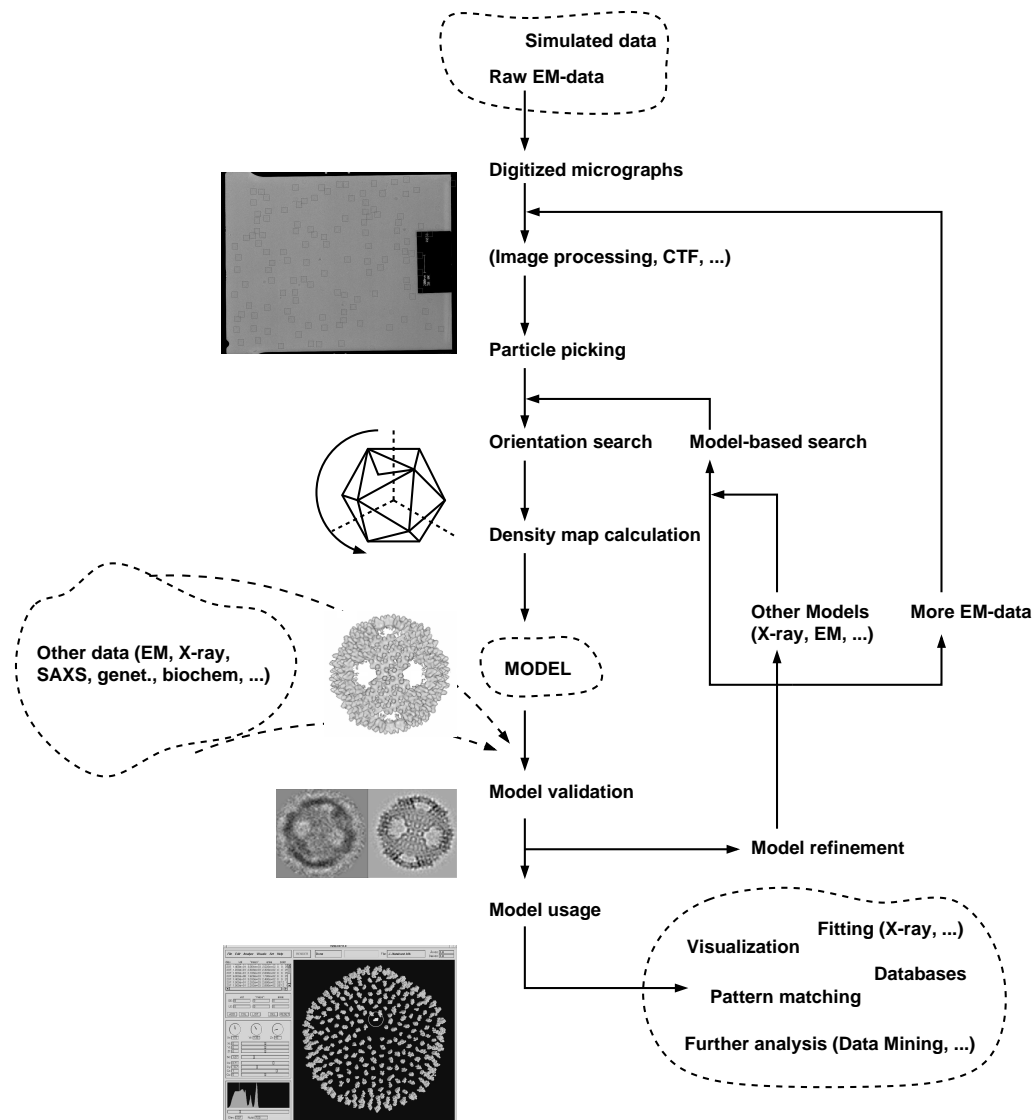
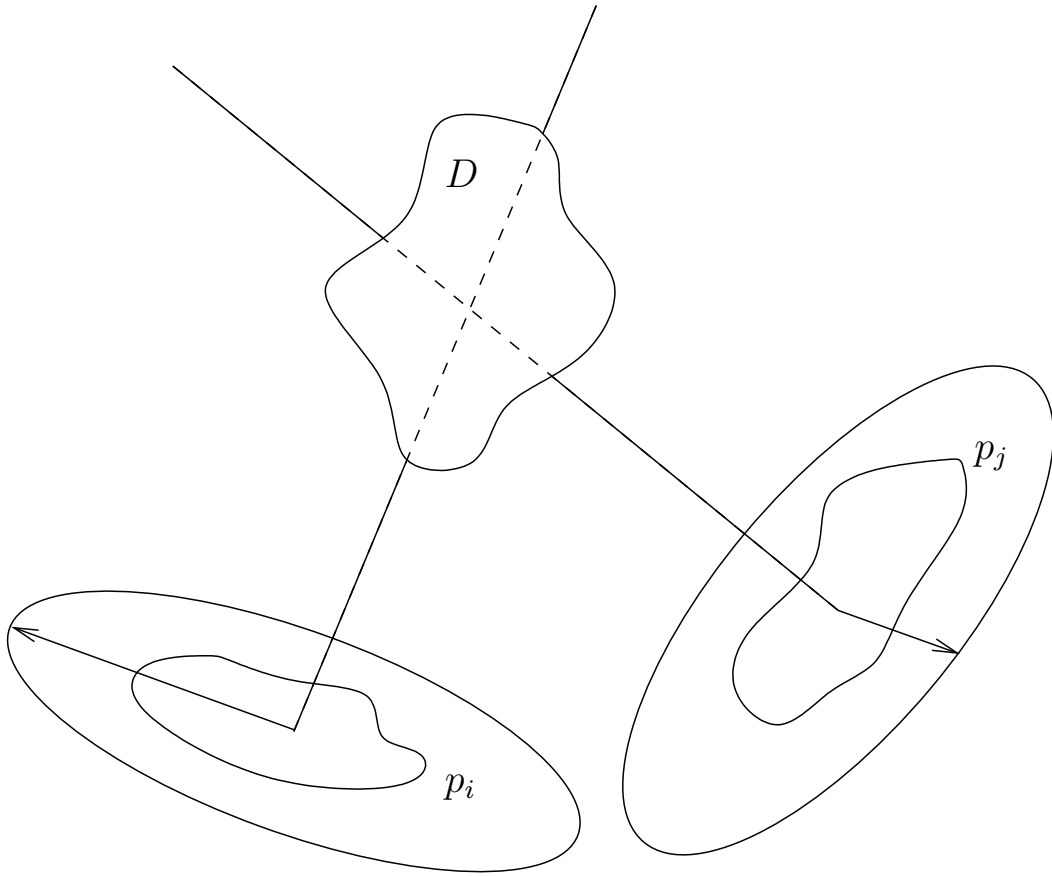


Figure 1: The reconstruction process.

Figure 2: Two projections of density D .

compare how well the given projections fit to the projections of the reconstructed density map. This kind of definition of good orientations suggests an Expectation Maximization-type procedure of repeatedly finding the best model for fixed orientations and the best orientations for a fixed model, see e.g. [3, 11, 20, 22, 30]. Due to the strong dependency on the reconstruction method, it is not easy to say analytically much (even whether it converges) about this approach in general. In practice, this approach to orientation search works successfully if there is an approximate density map of the particle available to be used as an initial model.

The orientations can be determined also by *common lines* [2]: Let p_i and p_j be projections of a density map D onto planes corresponding to orientations o_i and o_j , respectively; see Figure 2. All one-dimensional projections of D onto a line passing through the origin in the plane corresponding to the orientation o_i (o_j) can be computed from the projection p_i (p_j); this collection of projections of p_i (p_j) is also called the *sinogram* of p_i (p_j). As the two planes intersect, there is a line for which the projections of p_i and p_j agree. This line (which actually is a vector since the one dimensional projections are oriented, too) is

called the *common line* of p_i and p_j ; Figure 3.

If the projections are noiseless then already the pairwise common lines of three projections determine the relative orientations of the projections in three-dimensional space uniquely (except for the handedness) provided that the possible symmetries of the particle are taken into account. Furthermore, this can be computed by only few arithmetic and trigonometric operations [29].

However, the projections produced by the electron microscope are extremely noisy and so it is highly unlikely that two projections have one-dimensional projections that are equal. In this case it would be natural to try to find the best possible approximate common lines, i.e., a pair of approximately equal rows from the sinograms for the two projections. Several heuristics for the problem have been proposed [4, 5, 9, 10, 16, 23, 27, 28, 29]. However, they usually assume that the density map under reconstruction is highly symmetric which radically improves the signal-to-noise ratio. In Section 3 we partially justify the use of heuristics by showing that many variants of the orientation search problem are computationally very difficult.

3 The Complexity of Orientation Search

In this section we show that finding good orientations using common lines is computationally very difficult in general but it has some efficiently solvable special cases. The results are described in three phases: First, we consider the decision versions of the orientation search problem. Second, we study the approximability of several optimization variants. Finally, we examine the parameterized (in)tractability of the problem.

We would like to point out that some of the results are partially similar to the results of Hallett and Lagergren [17] for their problem CORE-CLIQUE that models the problem of finding functionally analogous genes. However, our problem of finding good orientations based on common lines differs from the problem of finding functionally analogous genes, e.g., by its geometric nature and by its very different application domain. Furthermore, we provide relevant positive results for finding functionally analogous genes: we describe an approximation algorithm with guaranteed approximation ratio of $2\beta(1 - o(1))$, if the distances between genes adhere to the triangle inequality within a factor β .

3.1 Decision Complexity

As mentioned in Section 2, the pairwise common lines cannot be detected reliably when the projections are very noisy. A natural relaxation is to allow several common line candidates for each pair of projections. In this section

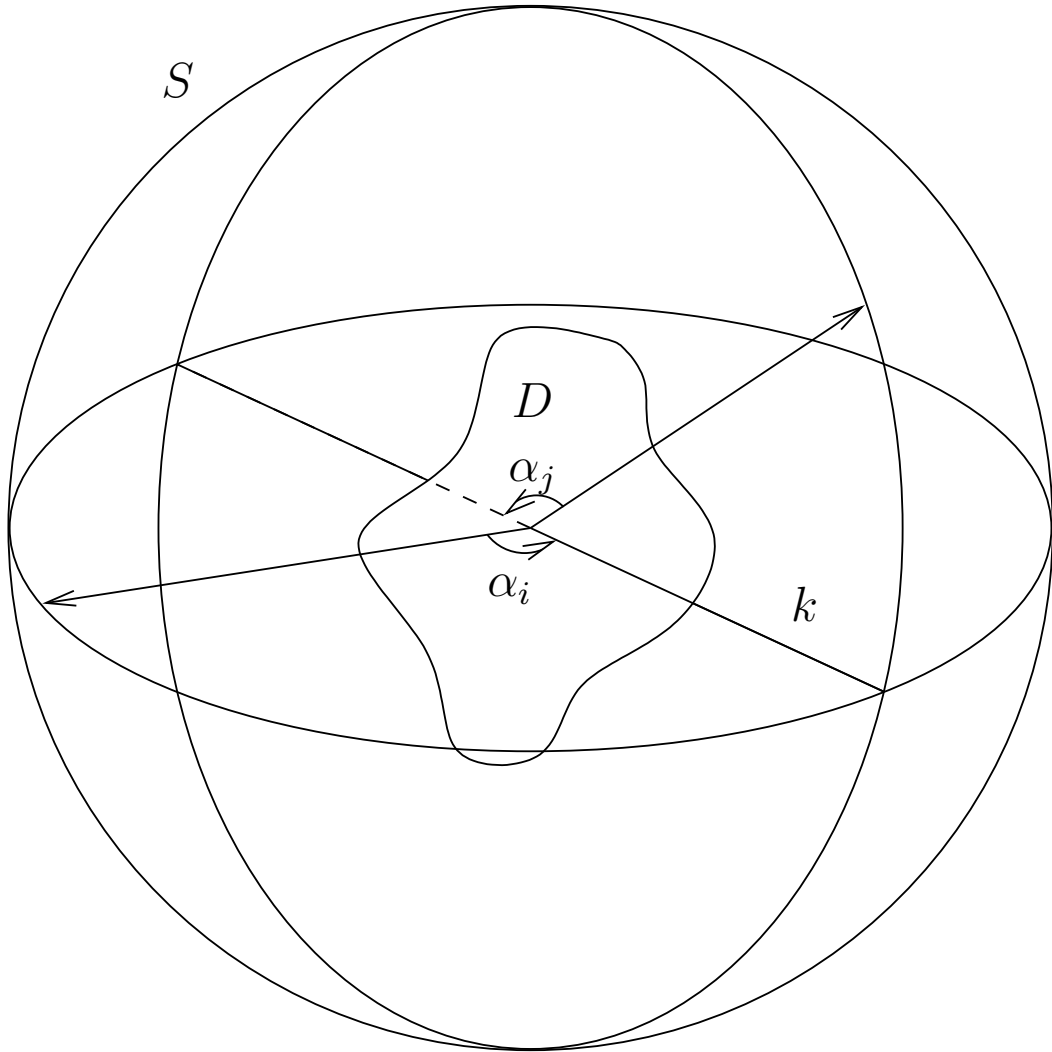


Figure 3: Two projection directions presented as great circles and their common line k specified with the rotation angles α_i and α_j in the internal coordinate systems of the two circles.

we study the problem of deciding whether there exist common lines in given sets of pairwise common lines that determine consistent orientations. We show that some formulations are *NP*-complete in general but there are nontrivial special cases that are solvable in nondeterministic logarithmic space. (For further information about computational complexity and complexity classes of decision problems, see e.g. [26].)

The common lines-based orientation search problem can be modeled at a high level as the problem of finding an n -clique from an n, m -partite graph $G = (V_1, \dots, V_n, E)$, i.e., a graph consisting independent sets V_1, \dots, V_n of size m .

Problem 1 (*n*-clique in an n, m -partite graph) *Given an n, m -partite graph $G = (V_1, \dots, V_n, E)$, decide whether there is an n -clique in G .*

Problem 1 can be interpreted as the orientation search problem in the following way: each group V_i describes the possible orientations of the projection p_i and each edge connecting two oriented projections says that the projections in the corresponding orientations are consistent with each other.

On one hand already three different orientations for each projection can make the problem *NP*-complete:

Theorem 1 *Problem 1 is NP-complete if $m \geq 3$.*

Proof. Clearly, the problem is in *NP* since one can check in polynomial time in $|G|$ whether a given subset of the vertices of G forms an n -clique.

We show the *NP*-hardness of Problem 1 by reduction from the graph k -colorability problem:

Problem 2 (graph k -colorability [26]) *Given a graph $G = (V, E)$ and a positive integer k , decide whether G is k -colorable, i.e., whether there is a mapping $f : V \rightarrow \{1, \dots, k\}$ such that if $\{u, v\} \in E$ then $f(u) \neq f(v)$.*

Let $G' = (V', E')$ be the graph that we would like to color with k colors. The polynomial-time reduction to a corresponding instance $G = (V_1, \dots, V_n, E)$ of Problem 1 is as follows. For each vertex $i \in V'$ there is a group V_i consisting of k vertices v_i^1, \dots, v_i^k . Each vertex in V_i corresponds to one coloring of the vertex $i \in V'$. There is an edge $\{v_i, v_j\} \in E, v_i \in V_i, v_j \in V_j, i \neq j$, if and only if $\{i, j\} \notin E'$ or v_i and v_j are of different color.

Clearly, the graph $G' = (V', E')$ is k -colorable if and only if there is an n -clique in the corresponding n, k -partite graph $G = (V_1, \dots, V_n, E)$. The members of groups V_i that correspond to a coloring form an n -clique in G . \square

On the other hand the problem can be solved in nondeterministic logarithmic space if the number of orientations for each projection is at most two:

Theorem 2 *Problem 1 is NL -complete if $m \leq 2$.*

Proof. The problem is in NL since it can be reduced in logarithmic space to the m -satisfiability problem with $m \leq 2$ that is an NL -complete problem:

Problem 3 (m -satisfiability [26]) *Given a set U of boolean variables and a set C of clauses $c \in C, |c| \leq m$, decide whether there is a truth value assignment $f : U \rightarrow \{0, 1\}$ that satisfies all clauses in C , i.e., whether there is a truth value assignment f that sets at least one literal¹ true in each clause of C .*

Note first that any instance of the problem with $m \leq 2$ can be trivially reduced to the case with $m = 2$. The reduction from Problem 1 with $m = 2$ to Problem 3 with $m = 2$ is as follows. Let the instance of Problem 1 be $G = (V_1, \dots, V_n, E)$ and the instance of Problem 3 (U, C) . For each group $V_i = \{v_i^0, v_i^1\}$ there is a boolean variable u_i whose truth value assignments $u_i = 0$ and $u_i = 1$ correspond to vertices v_i^0 and v_i^1 , respectively. The set C contains a clause $u_i = (1 - a)^2 \vee u_j = (1 - b)^2$ if and only if $\{v_i^a, v_j^b\} \notin E$.

If there is a truth assignment f satisfying all clauses in C then the vertices corresponding to the truth value assignments form an n -clique V' in G : Assume contrary that the truth value assignment f satisfies all clauses in C but the corresponding set V' of n vertices does not form an n -clique. Then there are at least two vertices v_i^a and v_j^b in V' such that $\{v_i^a, v_j^b\} \notin E$. But then C contains a clause $u_i = (1 - a)^2 \vee u_j = (1 - b)^2$ which the truth value assignment f does not satisfy. If no truth value assignment f satisfies all clauses in C then in any set V' of n vertices there are at least two vertices v_i^a and v_j^b such that $\{v_i^a, v_j^b\} \notin E$.

Thus, the graph G contains an n -clique if and only if there is a truth value assignment f that satisfies all clauses in C .

The problem is also NL -hard since Problem 3 with $m = 2$ can be reduced to it in logarithmic time in a similar way. \square

The formulation of the orientation search problem as Problem 1 seems to miss some of the geometric nature of the problem. As a first step toward the final formulation, let us consider the problem of finding a constrained line arrangement, the constraint being that any two lines of the arrangement are allowed to intersect only at a given set of points, each such set being of size $\leq l$:

Problem 4 (l -constrained line arrangement) *Given sets $P_{ij} \subset \mathbb{R}^2, |P_{ij}| \leq l, 1 \leq i < j \leq n$, decide whether there exist lines L_1, \dots, L_n in \mathbb{R}^2 such that L_i and L_j intersect only at some $p \in P_{ij}$ for all $1 \leq i < j \leq n$.*

¹Recall that literals are just boolean formulas of type $x = 0$ and $x = 1$ where x is a variable.

This problem has some interest of its own since line arrangements are one of the central concepts in computational and discrete geometry [13, 24]. If we require that the lines are in general position, i.e., that they are not parallel nor they intersect in same points, then we get the following hardness result:

Theorem 3 *Problem 4 is NP-complete if $l \geq 9$.*

Proof. The problem is in NP for all $l \geq 0$ since it can be checked in polynomial time whether there are lines L_1, \dots, L_n such that L_i and L_j intersect at p_{ij} for each $1 \leq i < j \leq n$.

The NP-hardness of the problem can be shown by a polynomial-time reduction from Problem 1 as follows. Let $G = (V_1, \dots, V_n, E)$ be the instance of Problem 1. For each vertex $v_{i,a} \in V_i$ we have a line $L_{i,a}$. Set P_{ij} contains the intersection point of lines $L_{i,a}$ and $L_{j,b}$ if and only if $\{v_{i,a}, v_{j,b}\} \in E$. We can use this reduction if we are able to find nm lines on plane in *general position* (for discussion on what being in general position means, see [24]). Actually, it is sufficient to require that

1. no two lines are parallel,
2. no three lines intersect in the same point, and
3. if $p_{ij_1} \in P_{ij_1}$, $p_{ij_2} \in P_{ij_2}$ and $p_{ij_3} \in P_{ij_3}$ are on same line then this line is one of the lines $L_{i,a}$.

Non-vertical lines $y = gx + h$ can be mapped to points $(g, h) \in \mathbb{R}^2$ and vice versa. The nm lines can be generated by considering the pairs $(g, h) \in \mathbb{N}^2$ of positive integers in lexicographical order \prec : $(g_1, h_1) \prec (g_2, h_2)$ if and only if $g_1 < g_2 \vee (g_1 = g_2 \wedge h_1 < h_2)$; and choosing some points (g, h) according to rules that are equivalent to the above rules for lines. The rules for choosing the points are:

1. each chosen point has a unique first coordinate g ; we call g the *column index* of the point,
2. no line passes through three chosen points, and
3. three lines, each passing through two chosen points, can intersect in the same point only if that point is chosen, too.

We still have to show that it is sufficient to consider only a polynomial number of points in \mathbb{N}^2 in order to find nm points that satisfy the given requirements. Let the number of chosen points at certain stage of the construction to be k with one point chosen from each column $0, \dots, k-1$. Then the maximum number of the points we have to consider at column k before finding a feasible point can be bounded above polynomially in n and m as follows:

- Exactly $\binom{k}{2}$ lines can be drawn passing through at least two chosen points. These lines make at most $\binom{k}{2}$ points in the column k infeasible.
- Two points on a plane span a line uniquely. Any four chosen points span two different lines L'_i and L'_j and there are exactly $\binom{k}{4}$ such pairs of lines. Each of the other $k - 4$ chosen points can span at most one line with the points in the column k that passes through the intersection point of the lines L'_i and L'_j . Thus, the number of points in the column k that are infeasible due to this is at most $(k - 4) \binom{k}{4}$.

Thus, the number of points in column k that have to be considered before finding the first point that does not violate our selection rules and hence can be chosen as the $k + 1$:st point is at most

$$\binom{k}{2} + (k - 4) \binom{k}{4}$$

which is clearly polynomial in nm when $k \leq nm$. □

The result can be slightly improved if we relax the general position requirement used in Theorem 3, e.g., if we allow also parallel lines in the arrangement:

Theorem 4 *Problem 4 is NP-complete if $l \geq 6$.*

Proof. The problem is in NP as noted in the proof of Theorem 3.

The NP-hardness of the problem can be shown by reduction from Problem 3 as follows. Given an instance (C, U) of the m -satisfiability problem, we construct point sets P_{ij} for $1 \leq i \leq |U|$ and $1 \leq j \leq |C|$. This is done by representing the variables and clauses by suitable line arrangements and constraining their intersection points. Each boolean variable $u_i \in U$ is represented by two vertical lines L_i^0 and L_i^1 representing the truth value assignments $u_i = 0$ and $u_i = 1$, respectively. Each clause $c_j \in C$ is represented by $|c_j|$ horizontal lines $L_{j,1}, \dots, L_{j,|c_j|}$. The intersection point of lines $L_{i,a}$ and $L_{j,b}$ corresponding to the truth value assignment $u_i = a$ and the b th literal in the clause c_j is in P_{ij} if and only if the truth value assignment $u_i = a$ does not falsify the b th literal in c_j which fixes sets P_{ij} . These lines are placed on plane in such way that all vertical lines have different horizontal coordinates and all horizontal lines have different vertical coordinates.

Without loss of generality, we assume that $|U| > m$ and $|C| > 2$. This ensures that all lines that are spanned by the points in sets P_{ij} and correspond to the clauses must be horizontal and all lines that correspond to the variables must be vertical in any feasible line arrangement corresponding to a satisfying truth assignment.

If there is a satisfying truth assignment f for the set C of clauses then the lines of the corresponding line arrangement intersect in the allowed points

that belong to the sets P_{ij} . If there are lines intersecting only at the allowed points then the vertical lines uniquely determine a truth value assignment f that satisfies all clauses in C .

Thus, the lines can be arranged on plane in such way that they intersect only at allowed intersection points in sets P_{ij} if and only if there is a truth value assignment satisfying all clauses in C . Furthermore, if the size of the largest clause is m then the size of the largest set P_{ij} is at most $2m = l$. As Problem 3 is NP -complete when $m \geq 3$, Problem 4 is NP -complete when $l \geq 6$. \square

However, the orientation search is not about arranging lines on the plane but great circles on the (unit) sphere $S = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$ as the orientations and the great circles are obviously in one-to-one correspondence. Thus, we should study the great circle arrangements:

Problem 5 (l -constrained great circle arrangement) *Given sets $P_{ij} \subset S_+ = \{(x, y, z) \in S : z \geq 0\}$, $|P_{ij}| \leq l$, $1 \leq i < j \leq n$, decide whether there exist great circles C_1, \dots, C_n on S such that C_i and C_j intersect on S_+ only at some $p \in P_{ij}$ for all $1 \leq i < j \leq n$.*

It can be shown that the line arrangements and great circle arrangements are equivalent through the stereographic projection [13]:

Theorem 5 *Problem 5 is as difficult as Problem 4.*

Proof. Great circles on a sphere can be mapped to lines on a plane by the central projection and lines on a plane to great circles on a sphere by its inverse [7]. \square

Still, our problem formulation is lacking some of the important ingredients of the orientation search problem: it is not possible to express at this stage of the orientation search the common line candidates by giving the allowed pairwise intersection points on the sphere S , i.e., in some globally fixed coordinate system. Rather, one can represent a common line only in the internal coordinates of the two great circles that correspond to the two projections intersecting. Each coordinate is in fact an angle giving the rotation angle of the common line on the projection as depicted in Figure 3. Hence the representation is a pair of angles:

Problem 6 (locally l -constrained great circle arrangement on sphere) *Given sets $P_{ij} \subset [0, 2\pi) \times [0, 2\pi)$, $|P_{ij}| \leq l$, $1 \leq i < j \leq n$, decide whether there exist great circles C_1, \dots, C_n on S such that C_i and C_j intersect only at some $p \in P_{ij}$ for all $1 \leq i < j \leq n$, where p defines the angles of the common line on C_i and C_j .*

Also this problem can be shown to be equally difficult to decide:

Theorem 6 *Problem 6 is NP-complete if $l \geq 6$.*

Proof. The problem is in NP since it is possible to check in polynomial time in the total number of possible local intersection points whether a given set of local intersection points is realizable.

The NP-hardness of the problem can be obtained from the proofs of Theorem 4 and Theorem 5. Indeed, all great circles corresponding to the horizontal lines in the corresponding line arrangement are forced to be parallel by their common intersection point. Similarly, all great circles corresponding to the vertical lines are forced to be parallel by their common intersection point. \square

Thus, deciding whether there exist consistent orientations seems to be difficult in general.

3.2 Approximability

As finding a consistent orientation for the projections is by the results of Section 3.1 difficult, we should consider also orientations that may determine orientations only for a large subset of the projections or resort to common lines that are as good as possible.

A simple approach to consider consistent orientations for large subsets of the projections is to look for large cliques in the n, m -partite graph $G = (V_1, \dots, V_n, E)$ instead of exactly n -cliques. In the world of orientations this means that instead of finding consistent orientations for all projections we look for consistent orientations for as many projections as we are able to and neglect the other projections.

Containing a clique is just one example of a property a graph can have. Also other graph properties might be useful. Thus we can formulate the problem in a rather general form as follows:

Problem 7 (Maximum subgraph with property P in an n, m -partite graph)

Given an n, m -partite graph $G = (V_1, \dots, V_n, E)$, find the largest $V' \subseteq V_1 \cup \dots \cup V_n$ such that the induced subgraph satisfies the property P and $|V' \cap V_i| \leq 1$ for all $1 \leq i \leq n$.

This resembles the following fundamental graph problem in combinatorial optimization and approximation algorithms:

Problem 8 (Maximum subgraph with property P [1]) *Given a graph $G = (V, E)$, find the largest $V' \subseteq V$ such that the induced subgraph satisfies the property P .*

It is not very difficult to see that the two problems are equivalent:

Theorem 7 *Problem 7 is as difficult as Problem 8.*

Proof. On the one hand, Problem 7 is a special case of Problem 8 with a restricted graph structure and with the additional condition $|V' \cap V_i| \leq 1$ for all i which can be included in the property P . On the other hand, Problem 8 is a special case of Problem 7 with singleton groups $V_1, \dots, V_{|C|}$. \square

Problem 8 is very difficult w.r.t. several properties [1]. By Theorem 7, these results generalize to Problem 7. Hence, for example, finding the maximum clique from the n, m -partite graph cannot be approximated within ratio $n^{1-\epsilon}$ for any fixed $\epsilon > 0$ [19]. Note that the approximation ratio n can be achieved trivially by choosing any of the vertices in G which is always a clique of size 1.

In practice the techniques for finding common lines or common line candidates actually evaluate all potential common lines of two projections (that is, all relative orientations of the two projections with respect to each other are in effect considered) and give them a score which typically is the distance between the two sinogram rows corresponding to potential common line. Thus, we could assume that there is always at least one feasible solution and study the following problem:

Problem 9 (Minimum weight n -clique in a complete n, m -partite graph)

Given a complete n, m -partite graph $G = (V_1, \dots, V_n, E)$ and a weight function $w : E \rightarrow \mathbb{N}$, find $V' \subset V_1 \cup \dots \cup V_n$ such that the weight $\sum_{u,v \in V', u \neq v} w(\{u, v\})$ is minimized and $|V' \cap V_i| \leq 1$ for all $1 \leq i \leq n$.

Unfortunately, it turns out that in this case the situation is extremely bad:

Theorem 8 *Problem 9 with $m \geq 3$ is not polynomial-time approximable within 2^{n^k} for any fixed $k > 0$ if $P \neq NP$.*

Proof. If Problem 9 were approximable within 2^{n^k} for some fixed $k > 0$ then the NP -complete Problem 1 could be solved in polynomial time by using the following weight function for the edges:

$$w(e) = \begin{cases} 2^{n^k} & \text{if } e \in E \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the problem is not approximable within 2^{n^k} in polynomial time provided that $P \neq NP$. \square

When there are only two vertices in each group the problem admits a constant factor approximation ratio but no better:

Theorem 9 *Problem 9 is APX-complete if $m = 2$.*

Proof. This can be shown by an approximation-preserving reduction from and to the minimum weight 2-satisfiability problem that is known to be *APX*-complete:

Problem 10 (Minimum weight 2-satisfiability [1]) *Given a set U of boolean variables, a set C of clauses $c \in C, |c| \leq 2$ and a weight function $w : C \rightarrow \mathbb{N}$, find the truth value assignment $f : U \rightarrow \{0, 1\}$ that minimizes the sum of the weights of unsatisfied clauses, i.e.,*

$$\sum_{U \text{ does not satisfy } c \in C} w(c).$$

The reduction from Problem 10 to Problem 9 with $m = 2$ is very similar to the reduction in the proof of Theorem 2. Each boolean variable u_i is represented by a two-set $V_i = \{v_i^0, v_i^1\}$ corresponding to truth value assignments $u_i = 0$ and $u_i = 1$, respectively. By definition of Problem 9, the graph $G = (V_1, \dots, V_n, E)$ is complete, i.e., $E = \{\{u, v\} : u \in V_i, v \in V_j, 1 \leq i < j \leq n\}$. The weight of the edge $e = \{v_i^a, v_j^b\} \in E$ is zero if there is a clause $u_i = (1 - a)^2 \vee u_j = (1 - b)^2$ in C and $w(e)$ otherwise.

Thus, the weight of the n -clique V' in G equals to the weight of the clauses that are not satisfied by the truth value assignment corresponding to the n -clique determined by V' . That is, Problem 9 with $m = 2$ is at least as difficult as Problem 10.

Problem 9 with $m = 2$ can be reduced in polynomial time to Problem 10 in a similar way. For each vertex set V_i in G there is a boolean variable u_i and the vertices $v_i^0, v_i^1 \in V_i$ correspond to the two truth value assignments of u_i . For each edge $e = \{v_i^a, v_j^b\}$ in E there is a clause $u_i = (1 - a)^2 \vee u_j = (1 - b)^2$ with weight $w(e)$.

The weight of the clauses that the truth value assignment f does not satisfy is equal to the weight of the corresponding n -clique in G . That is, Problem 9 with $m = 2$ is at most as difficult as Problem 10.

Thus, Problem 9 with $m = 2$ is *APX*-complete, as claimed. \square

An easier variant of Problem 9 is the case where the edge weights admit the triangle inequality within a factor β , i.e., for all edges $\{t, u\}$, $\{t, v\}$ and $\{u, v\}$ in E it holds

$$w(\{t, u\}) \leq \beta(w(\{t, v\}) + w(\{u, v\})).$$

A good approximation of the minimum weight n -clique in G can be found by finding the minimum weight n -star that contains one vertex from each group V_i . The method is described by Algorithm 1.

Algorithm 1 gives constant-factor approximation guarantees and the approximation is stable (for details on approximation stability, see [6]):

Algorithm 1 A constant-factor approximation algorithm for finding the minimum weight n -clique from a weighted graph.

```

1: function MINIMUM-WEIGHT-STAR( $G, w$ )
2:    $W_{\min} \leftarrow \infty$ 
3:   for  $i = 1, \dots, n$  do
4:     for all  $v \in V_i$  do
5:        $W \leftarrow 0$ 
6:       for  $j = 1, \dots, i-1, i+1, \dots, n$  do
7:          $W \leftarrow W + \min_{u \in V_j} \{w(\{u, v\})\}$ 
8:       end for
9:       if  $W < W_{\min}$  then
10:         $W_{\min} \leftarrow W$ 
11:         $v_{\min} \leftarrow v$ 
12:       end if
13:     end for
14:   end for
15:    $V' \leftarrow \emptyset$ 
16:   for  $j = 1, \dots, n$  do
17:      $V' \leftarrow V' \cup \{\arg \min_{u \in V_j} \{w(\{u, v_{\min}\})\}\}$ 
18:   end for
19:   return  $(V', E' = \{e \in E : e \subseteq V'\})$ 
20: end function

```

Theorem 10 *Problem 9 is polynomial-time approximable within $2\beta(1 - o(1))$ by Algorithm 1 if the edge weights satisfy the triangle inequality within factor β .*

Proof. Let $G' = (V', E')$ be the n -clique found from the n, m -partite complete graph G by Algorithm 1 and let $OPT(G)$ be the minimum weight n -clique in G .

The weight of G' can be bounded above as follows. We distribute the weight of the solution G' to its vertices:

$$w(v) = \sum_{e \in E', v \in e} w(e)/2.$$

The weight of the lightest vertex in G' , the vertex v_{\min} , is

$$w(v_{\min}) = \sum_{e \in E', v_{\min} \in e} \frac{w(e)}{2} \leq \frac{n-1}{2n(n-1)/2} OPT(G) = \frac{1}{n} OPT(G).$$

For each edge $\{u, v\} \in E'$ such that $v_{\min} \notin \{u, v\}$, holds

$$w(\{u, v\}) \leq \beta [w(\{u, v_{\min}\}) + w(\{v, v_{\min}\})]$$

by the assumption. Thus, the sum of the weights of the other vertices in V'

$$\begin{aligned}
\sum_{v \in V', v \neq v_{\min}} w(v) &= \sum_{v \in V', v \neq v_{\min}} \sum_{e \in E', v \in e} \frac{w(e)}{2} \\
&\leq \sum_{v \in V', v \neq v_{\min}} \sum_{e \in E', v \in e} \frac{\beta [w(\{u, v_{\min}\}) + w(\{v, v_{\min}\})]}{2} \\
&= (n-1) 2\beta w(v_{\min}) = \frac{2\beta(n-1)}{n} OPT(G) \\
&= 2\beta \left(1 - \frac{1}{n}\right) OPT(G).
\end{aligned}$$

Combining these two upper bounds we get

$$w(G') \leq \frac{1}{n} OPT(G) + 2\beta \left(1 - \frac{1}{n}\right) OPT(G) = 2\beta(1 - o(1)) OPT(G).$$

Thus, Algorithm 1 guarantees the approximation factor $2\beta(1 - o(1))$ when w satisfies the triangle inequality within a factor β . \square

This algorithm might not be applicable in orientation search as there seems to be little hope of finding distance functions (used in selecting the best common lines) satisfying even the relaxed triangle inequality for the noisy projections. However, in the case of finding functionally analogous genes this is possible since many distance functions between sequences are metric. Thus, the algorithm seems to be very promising for that task.

A very natural relaxation of the original problem is to allow small changes to common line candidates to make the orientations consistent:

Problem 11 (Minimum error l -constrained line arrangement) *Given sets $P_{ij} \subset \mathbb{R}^2$, $|P_{ij}| \leq l$, $1 \leq i < j \leq n$, find lines L_1, \dots, L_n in \mathbb{R}^2 that minimize the sum of distances $\min_{p_{ij} \in P_{ij}} |p_{ij} - \hat{p}_{ij}|^q$ where \hat{p}_{ij} is the actual intersection point of lines L_i and L_j and $q > 0$.*

Unfortunately also this variant of the problem is very difficult:

Theorem 11 *Problem 11 with $l \geq 6$ is not polynomial-time approximable within 2^{n^k} for any fixed $k > 0$ if $P \neq NP$.*

Proof. If Problem 11 would be polynomial-time approximable within 2^{n^k} for some fixed $k > 0$ then Problem 4 could be solved in polynomial time since there are lines intersecting at the allowed points if and only if the minimum error line arrangement has error zero. \square

3.3 Parameterized Complexity

Even if the problem is *NP*-hard, it might be solvable in practice if the *NP*-hardness is caused by some properties of the inputs that do not occur in practice. For example, one might be interested only vertex covers of size at most k . Deciding whether there is vertex cover of size at most k in a graph of n vertices can be solved in time $O(n^k)$. However, if k is, e.g., 40 and n is very large then this time complexity is unacceptable. Instead, we would like to have time complexity of form $O(n^c)$ for some reasonably small c .

Formally a *parameterized decision problem* is a set $D \subseteq \Sigma^* \times \mathbb{N}$ where Σ is a finite alphabet. A parameterized decision problem D is *fixed-parameter tractable* if for each $(x, k) \in \Sigma^* \times \mathbb{N}$ it can be decided whether (x, k) is in D in time $f(k) |x|^{O(1)}$ where $f : \mathbb{N} \rightarrow \mathbb{N}$ is an arbitrary function. Parameterized complexity classes form a hierarchy similar to the polynomial hierarchy:

$$FPT \subseteq W[1] \subseteq W[1] \subseteq \dots \subseteq W[SAT] \subseteq W[P].$$

All inclusions between the classes are believed to be proper. All problems outside the class *FPT* are called *fixed-parameter intractable*.

A parameterized problem D reduces to a parameterized problem D' if there exist functions $f, g : \mathbb{N} \rightarrow \mathbb{N}$ and $h : D \rightarrow D'$ such that $h(x, k)$ is computable in time $f(k) |x|^{O(1)}$ for each instance in $\Sigma \times \mathbb{N}$, and $(x, k) \in D$ if and only if $(h(x), g(k)) \in D'$. Such a reduction is called a *standard parameterized m -reduction*. (For further details on parameterized complexity, see [12].)

For the orientation search problem there is a natural parameterization: the number of projections can be bounded by a constant. Thus, Problem 1 can be turned into the following parameterized problem:

Problem 12 (k -clique in k, m -partite graph) *Given a k, m -partite graph $G = (V_1, \dots, V_k, E)$ and a natural number k , decide whether there is a k -clique in G .*

The intuition behind this formulation of being interesting is that if we would be able to orient a few representative projections very well then the risk that the orientations found for the other projections based on those well-oriented representative projections would be incorrect could be small enough. Thus, there would be good chances to reconstruct an accurate density map based on the found orientations. Unfortunately, also this formulation is fixed-parameter intractable:

Theorem 12 *Problem 12 is $W[1]$ -complete.*

Proof. Let us first show that the k, m -satisfiability is $W[1]$ -hard:

Problem 13 (k, m -satisfiability) *Given a set U of boolean variables and a set $C, |C| = k$, of clauses $c \in C, |c| \leq m$, decide whether there is a truth value assignment $f : U \rightarrow \{0, 1\}$ that satisfies all clauses in C .*

Lemma 1 *Problem 13 is $W[1]$ -complete.*

Proof. The problem is shown to be $W[1]$ -hard by a parameterized reduction from the short nondeterministic Turing machine computation problem which is known to be $W[1]$ -complete.

Problem 14 (Short nondeterministic Turing machine computation [12])

Given a nondeterministic Turing machine M , input string x and a natural number k , decide whether there is a computation of M that accepts the string x in at most k steps.

It can be verified that the reduction used in Cook's Theorem (see e.g. [18]) is a parameterized reduction. Thus, it can be used also here to show that Problem 14 reduces to Problem 13.

Problem 13 can be shown to be in $W[1]$ by reduction to Problem 14. \square

Problem 12 is $W[1]$ -hard by a reduction from Problem 13 as follows. For each clause $c_i \in C$ there is a group V_i consisting of vertices corresponding to the literals in c_i . There is an edge between $v_{i,a} \in V_i$ and $v_{j,b} \in V_j$ if and only if $i \neq j$ and the corresponding literals can be satisfied simultaneously.

Problem 12 can be shown to be in $W[1]$ by a reduction to Problem 14. \square

4 Conclusions

We have shown that some approaches for determining orientations for noisy projections of identical particles are computationally very difficult, namely NP -complete, inapproximable and fixed-parameter intractable. These results justify (to some extent) the heuristic approaches widely used in practice.

On the bright side, we have been able to detect some polynomial-time solvable special cases. Also, we have described an approximation algorithm that achieves the approximation ratio $2\beta(1 - o(1))$ if the instance admits the triangle inequality within a factor β . It has promising applications in search for functionally analogous genes.

As a future work we wish to study the usability of current state of art in heuristic search to find reasonable orientations in practice. This is very challenging due to the enormous size of the search space. Another goal is to analyze the complexity of other approaches for determining the orientations for the projections.

References

- [1] G. Ausiello, P. Crescenzi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Springer-Verlag, 1999.
- [2] T. S. Baker, N. H. Olson, and S. D. Fuller. Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral. *Microbiology and Molecular Biology Reviews*, 63(4):862–922, 1999.
- [3] Timothy S. Baker and R. Holland Cheng. A model-based approach for determining orientations of biological macromolecules imaged by cryoelectron microscopy. *Journal of Structural Biology*, 116:120–130, 1996.
- [4] Pier Luigi Bellon, Francesca Cantele, and Salvatore Lanzavecchia. Correspondence analysis of sinogram lines. Sinogram trajectories in factor space replace raw images in the orientation of projections of macromolecular assemblies. *Ultramicroscopy*, 87:187–197, 2001.
- [5] Pier Luigi Bellon, Salvatore Lanzavecchia, and Vladimiro Scatturin. A two exposures technique of electron tomography from projections with random orientation and a *quasi*-Boolean angular reconstitution. *Ultramicroscopy*, 72:177–186, 1998.
- [6] Hans-Joachim Böckenhauer, Juraj Hromkovič, Ralf Klasing, Sebastian Seibert, and Walter Unger. Towards the notion of stability of approximation for hard optimization tasks and the traveling salesman problem. *Theoretical Computer Science*, 185(1):3–24, 2002.
- [7] Jean-Daniel Boissonat and Mariette Yvinec. *Algorithmic Geometry*. Cambridge University Press, 1998.
- [8] J. M. Carazo, C. O. Sorzano, E. Rietzel, R. Schröder, and R. Marabini. Discrete tomography in electron microscopy. In Gabor T. Herman and Attila Kuba, editors, *Discrete Tomography: Foundations, Algorithms, and Applications*, Applied and Numerical Harmonic Analysis, chapter 18, pages 405–416. Birkhäuser, 1999.
- [9] José R. Castón, David M. Belnap, Alasdair C. Steven, and Benes L. Trus. A strategy for determining the orientation of refractory particles for reconstruction from cryo-electron micrographs with particular reference to round, smooth-surfaced, icosahedral viruses. *Journal of Structural Biology*, 125:209–215, 1999.
- [10] R.A. Crowther, D.J. DeRosier, and A. Klug. The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. *Proceedings of the Royal Society of London A*, 317:319–340, 1970.

- [11] Peter C. Doerschuk and John E. Johnson. *Ab initio* reconstruction and experimental design for cryo electron microscopy. *IEEE Transactions on Information Theory*, 46(5):1714–1729, 2000.
- [12] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Monographs in Computer Science. Springer-Verlag, 1999.
- [13] Herbert Edelsbrunner. *Algorithms in Combinatorial Geometry*, volume 10 of *EATCS Monographs on Theoretical Computer Science*. Springer-Verlag, 1987.
- [14] César Fernández and Gerhard Wider. TROSY in NMR studies of the structure and function of large biological macromolecules. *Current Opinion in Structural Biology*, 13:570–580, 2003.
- [15] Joachim Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press, 1996.
- [16] S. D. Fuller, S. J. Butcher, R. H. Cheng, and T. S. Baker. Three-dimensional reconstruction of icosahedral particles – the uncommon line. *Journal of Structural Biology*, 116:48–55, 1996.
- [17] Michael T. Hallett and Jens Lagergren. Hunting for functionally analogous genes. In Sanjiv Kapoor and Sanjiva Prasad, editors, *Foundations of Software Technology and Theoretical Computer Science*, volume 1974 of *Lecture Notes in Computer Science*, pages 465–476. Springer-Verlag, 2000.
- [18] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 2nd edition, 2001.
- [19] Johan Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Mathematica*, 182:105–142, 1999.
- [20] Yongchang Ji, Dan C. Marinescu, Wei Chang, and Timothy S. Baker. Orientation refinement of virus structures with unknown symmetry. In *Proceedings of the International Parallel and Distributed Processing Symposium*, pages 49–56. IEEE Computer Society, 2003.
- [21] Teemu Kivioja, Janne Ravantti, Anatoly Verkhovsky, Esko Ukkonen, and Dennis Bamford. Local average intensity-based method for identifying spherical particles in electron micrographs. *Journal of Structural Biology*, 131:126–134, 2000.
- [22] Cristopher J. Lanczycki, Calvin A. Johnson, Benes L. Trus, James F. Conway, Alasdair C. Steven, and Robert L. Martino. Parallel computing strategies for determining viral capsid structure by cryo-electron microscopy. *IEEE Computational Science & Engineering*, 5:76–91, 1998.

- [23] M. Lindahl. Strul – a method for 3D alignment of single-particle projections based on common line correlation in Fourier space. *Ultramicroscopy*, 87:165–175, 2001.
- [24] Jiří Matoušek. *Lectures on Discrete Geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer-Verlag, 2002.
- [25] William V. Nicholson and Robert M. Glaeser. Review: Automatic particle detection in electron microscopy. *Journal of Structural Biology*, 133:90–101, 2001.
- [26] Christos H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1995.
- [27] Pawel A. Penczek, Jun Zhu, and Joachim Frank. A common-lines based method for determining orientations for $N > 3$ particle projections simultaneously. *Ultramicroscopy*, 63:205–218, 1996.
- [28] Pamela A. Thuman-Commike and Wah Chiu. Improved common line-based icosahedral particle image orientation estimation algorithms. *Ultramicroscopy*, 68:231–255, 1997.
- [29] Marin van Heel. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy*, 21:11–124, 1987.
- [30] Zhye Yin, Yili Zheng, and Peter C. Doerschuk. An *ab Initio* algorithm for low-resolution 3-D reconstructions from cryoelectron microscopy images. *Journal of Structural Biology*, 133:132–142, 2001.