

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

Taneli Mielikäinen¹, Janne Ravantti², and Esko Ukkonen¹

¹ Department of Computer Science

² Institute of Biotechnology and Faculty of Biosciences

University of Helsinki, Finland

{`tmielika,ravantti,ukkonen`}@cs.Helsinki.FI

Abstract. In this paper 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 problem and show that while several variants of the problem are NP -hard and inapproximable, some restrictions are polynomial-time approximable within a constant factor or even solvable in logarithmic space. The negative complexity results give a partial justification for the heuristic methods used in the orientation search, and the positive complexity results have some positive implications also to a different problem of finding functionally analogous genes.

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 [1].

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 [2]. For example, there are many well-known viruses that do not seem to crystallize and are too large for NMR techniques.

A more flexible way to reconstruct density maps is offered by cryo-electron microscopy [1,3]. 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 [1]:

* A work supported by the Academy of Finland.

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. 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 mass distribution 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. 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.

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.

In this paper 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 [4].

The rest of this paper is organized as follows. In Section 2 the orientation search problem is described. Section 3 analyzes the computational complexity of the orientation search problem. The paper is concluded in Section 4.

Due to the page limitations the proofs of theorems and further details appear in the full version [5].

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

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 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. [6,7,8]. 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* [9]: Let p_i and p_j be projections of a density map D onto planes corresponding to orientations o_i and o_j , respectively. 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 .

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 [10].

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 of the two projections. Several heuristics for the problem have been proposed [3,10,11,12,13]. However, they usually assume that the density map under reconstruction is highly symmetric which radically improves the signal-to-noise ratio. In the next section we partially justify the use of heuristics by showing that many variants of the orientation search problem are computationally very difficult.

3 The Computational Complexity of Orientation Search Problem

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. First, we consider the decision versions of the orientation search problem. Second, we study the approximability of several optimization variants.

We would like to point out that some of the results are partially similar to the results of Hallett and Lagergren [4] 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 $\beta(2 - o(1))$, 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 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. Due to the page limitations the proofs and further details appear in the full version [5].

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 $m \geq 3$.*

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 $m = 2$.*

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 2 (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 $1 \leq i < j \leq n$.

This problem has some interest of its own since line arrangements are one of the central concepts in computational and discrete geometry [14]. 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 2 for lines L_i in general position is NP-complete if $l \geq 9$.*

The result can be slightly improved if we allow also parallel lines in the arrangement:

Theorem 4. *Problem 2 is NP-complete if $l \geq 6$.*

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 3 (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 [14]:

Theorem 5. *Problem 3 is as difficult as Problem 2.*

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 the common line candidates by giving the allowed pairwise intersection points on the sphere S . 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. Hence the representation is a pair of angles:

Problem 4 (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 4 is NP-complete.*

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 cover only a large subset of the projections or resort to common lines that are as good as possible.

A simple approach to allow errors in solutions is look for large cliques in the n, m -partite graph $G = (V_1, \dots, V_n, E)$ instead of exact 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 5 (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 6 (Maximum subgraph with property P [15]). 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 5 is as difficult as Problem 6.*

Thus, Problem 5 is very difficult w.r.t. several properties. For example, due to Theorem 7, finding the maximum clique from the n, m -partite graph cannot be approximated within ratio $n^{1-\epsilon}$ for any fixed $\epsilon > 0$ [16]. 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 produce distances between all possible intersections of two projections. Thus, we could assume that there is always at least one feasible solution and study the following problem:

Problem 7 (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 $V' \subseteq V_1 \cup \dots \cup V_n$ 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 7 with $m \geq 3$ is not polynomial-time approximable within 2^{n^k} for any fixed $k > 0$ if $P \neq NP$.*

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

Theorem 9. *Problem 7 is APX-complete if $m = 2$.*

An easier variant of Problem 7 is the case where the edge weights admit triangle inequality within a factor β , i.e., for all edges $\{t, u\}$, $\{t, v\}$ and $\{u, v\}$ it holds

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

A good approximation of the lightest clique can be found by finding the minimum weight star that contains one vertex from each group V_i . (The algorithm is described in the full version of this paper [5].) This gives constant-factor approximation guarantees and the approximation is stable (for details on approximation stability, see [17]):

Theorem 10. *Problem 7 is polynomial-time approximable within $\beta(2 - o(1))$ if the edge weights satisfy triangle inequality within factor β .*

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.

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

Problem 8 (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 $q > 0$.

Theorem 11. *Problem 8 with $l \geq 6$ is not polynomial-time approximable within 2^{n^k} for any fixed $k > 0$ if $P \neq NP$.*

4 Conclusions

In this paper we have shown that some approaches for determining orientations for noisy projections of identical particles are computationally very difficult, namely NP -complete and inapproximable. 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 $\beta(2 - o(1))$ 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. J. Frank, Three-Dimensional Electron Microscopy of Macromolecular Assemblies, Academic Press, 1996.

2. J. M. Carazo, C. O. Sorzano, E. Rietzel, R. Schröder, R. Marabini, Discrete tomography in electron microscopy, in: G. T. Herman, A. Kuba (Eds.), *Discrete Tomography: Foundations, Algorithms, and Applications*, Applied and Numerical Harmonic Analysis, Birkhäuser, 1999, Ch. 18, pp. 405–416.
3. R. Crowther, D. DeRosier, 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 (1970) 319–340.
4. M. T. Hallett, J. Lagergren, Hunting for functionally analogous genes, in: S. Kapoor, S. Prasad (Eds.), *Foundations of Software Technology and Theoretical Computer Science*, Vol. 1974 of *Lecture Notes in Computer Science*, Springer-Verlag, 2000, pp. 465–476.
5. T. Mielikäinen, J. Ravantti, E. Ukkonen, The computational complexity of orientation search problems in cryo-electron microscopy, Report C-2004-3, Department of Computer Science, University of Helsinki (2004).
6. P. C. Doerschuk, J. E. Johnson, *Ab initio* reconstruction and experimental design for cryo electron microscopy, *IEEE Transactions on Information Theory* 46 (5) (2000) 1714–1729.
7. Y. Ji, D. C. Marinescu, W. Chang, T. S. Baker, Orientation refinement of virus structures with unknown symmetry, in: *Proceedings of the International Parallel and Distributed Processing Symposium*, IEEE Computer Society, 2003, pp. 49–56.
8. C. J. Lanczycki, C. A. Johnson, B. L. Trus, J. F. Conway, A. C. Steven, R. L. Martino, Parallel computing strategies for determining viral capsid structure by cryo-electron microscopy, *IEEE Computational Science & Engineering* 5 (1998) 76–91.
9. T. S. Baker, N. H. Olson, S. D. Fuller, Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral, *Microbiology and Molecular Biology Reviews* 63 (4) (1999) 862–922.
10. M. van Heel, Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction, *Ultramicroscopy* 21 (1987) 11–124.
11. P. L. Bellon, F. Cantele, S. Lanzavecchia, Correspondence analysis of sinogram lines. Sinogram trajectories in factor space replace raw images in the orientation of projections of macromolecular assemblies, *Ultramicroscopy* 87 (2001) 187–197.
12. P. A. Penczek, J. Zhu, J. Frank, A common-lines based method for determining orientations for $N > 3$ particle projections simultaneously, *Ultramicroscopy* 63 (1996) 205–218.
13. P. A. Thuman-Commike, W. Chiu, Improved common line-based icosahedral particle image orientation estimation algorithms, *Ultramicroscopy* 68 (1997) 231–255.
14. H. Edelsbrunner, *Algorithms in Combinatorial Geometry*, Vol. 10 of *EATCS Monographs on Theoretical Computer Science*, Springer-Verlag, 1987.
15. G. Ausiello, P. Crescenzi, V. Kann, A. Marchetti-Spaccamela, M. Protasi, *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*, Springer-Verlag, 1999.
16. J. Hästad, Clique is hard to approximate within $n^{1-\epsilon}$, *Acta Mathematica* 182 (1999) 105–142.
17. H.-J. Böckenhauer, J. Hromkovič, R. Klasing, S. Seibert, W. Unger, Towards the notion of stability of approximation for hard optimization tasks and the traveling salesman problem, *Theoretical Computer Science* 185 (1) (2002) 3–24.