Parametric $k$-best alignment

Peter Huggins
Lane Center for Computational Biology
Carnegie Mellon University

Ruriko Yoshida
Department of Statistics
University of Kentucky

Abstract

Optimal sequence alignments depend heavily on alignment scoring parameters. Given input sequences, parametric alignment is the well-studied problem that asks for all possible optimal alignment summaries as parameters vary, as well as the optimality region of alignment scoring parameters which yield each optimal alignment. But biologically correct alignments might be suboptimal for all parameter choices. Thus we extend parametric alignment to parametric $k$-best alignment, which asks for all possible $k$-tuples of $k$-best alignment summaries $(s_1, s_2, \ldots, s_k)$, as well as the $k$-best optimality region of scoring parameters which make $s_1, s_2, \ldots, s_k$ the top $k$ summaries. By exploiting the integer-structure of alignment summaries, we show that, astonishingly, the complexity of parametric $k$-best alignment is only polynomial in $k$. Thus parametric $k$-best alignment is tractable, and can be applied at the whole-genome scale like parametric alignment.

Corresponding author:
name: Peter Huggins
email address: phuggins@andrew.cmu.edu
address:
1 Introduction

In pairwise sequence alignment, we are given a pair of homologous sequences \( \sigma_1, \sigma_2 \), and each alignment \( A \) of \( \sigma_1, \sigma_2 \) is endowed with an alignment summary \( s(A) \in \mathbb{Z}^d \) that records various features of \( A \), such as the number of mismatches and the number of spaces. Throughout we will assume the dimension \( d \) of alignment summaries is fixed, and that sequences are of length \( O(n) \). The score of alignment \( A \) is defined to be \( c \cdot s(A) \), where \( c \) is a fixed vector of alignment scoring parameters. A (global) optimal alignment is any alignment \( A \) which maximizes the alignment score. For most choices of alignment summary model, the optimal alignment summary can be computed in \( O(n^2) \) time by the Needleman–Wunsch (NW) algorithm [18], once the value \( c \) of the alignment scoring parameters is given.

The choice of \( c \) reflects relative frequencies of indels and different types of point mutations during sequence evolution. The optimal alignment is heavily dependent on the choice of \( c \), and yet in practice the “biologically correct” choice of \( c \) is not known. Given sequences \( \sigma_1, \sigma_2 \), the space of alignment scoring parameters partitions into optimality regions. Parameter values in the same optimality region give rise to the same optimal alignment summary. Parametric alignment [19] is the problem of determining all possible optimal alignment summaries that arise as \( c \) varies, and also computing the optimality region for each optimal summary.

Parametric alignment is a well-studied subject (see [8, 7, 10, 12, 13, 22]), and surprisingly tractable [19]: Pachter and Sturmfels proved that there are \( O(n^{d(d-1)/d+1}) \) optimality regions. In [13, 4], oracle-based methods for computing optimality regions are presented, which repeatedly run the NW algorithm with different choices of scoring parameters \( c \) to find new optimal alignments. Despite the nice \( O(n^{d(d-1)/d+1}) \) bound on the number of optimality regions, it has been speculated [8] that the required number of NW calls might be as high as \( \Theta(n^{d^2(d-1)/2d+1}) \). One purpose of this paper is to point out that, actually,

**Theorem 1.** Existing oracle-based methods for parametric alignment only use \( O(n^{d(d-1)/d+1}) \) calls to the NW algorithm.

Thus parametric alignment is much more tractable than previously thought. Nevertheless, one major shortcoming of parametric alignment is that it ignores nearly optimal alignments that are never optimal for any choice of scoring parameters. Nearly optimal alignments have been studied before (see [17, 21] and references within), but not in a parametric setting. Thus we propose parametric k-best alignment, which studies how the k-best alignment summaries vary with parameters. We consider two variants of the problem:

- **Ordered** parametric k-best alignment: Compute the collection of all ordered subsets of \( k \) distinct alignment summaries \( (s_1, \ldots, s_k) \) which can become the \( k \)-best summaries \( c \cdot s_1 > c \cdot s_2 > \ldots > c \cdot s_k > \ldots \) under some choice of \( c \). For each such subset \( (s_1, \ldots, s_k) \), find all \( c \) such that \( c \cdot s_1 \geq \ldots \geq c \cdot s_k \geq \ldots \) are the \( k \)-best summaries.
• *Unordered* parametric $k$-best alignment: Same problem, but the ordering of the $k$-best summaries $\{s_1, \ldots, s_k\}$ is ignored.

The output of parametric $k$-best alignment is a decomposition of the space of alignment scoring parameters into $k$-best optimality regions. All scoring parameters in a $k$-best optimality region yield the same list of $k$-best distinct alignment summaries.

Although parametric $k$-best alignment is a natural extension of parametric alignment, there are two major difficulties which have prevented its study:

- The structure of $k$-best optimality regions needs to be understood in order to systematically compute them, and
- naively, we might worry that the number of $k$-best optimality regions is exponential in $k$.

We first address the second point. Notice that the total number of subsets of $k$ alignment summaries grows exponentially in $k$. Indeed, if alignment summaries were arbitrary real-valued points, then the number of $k$-best optimality regions could be exponential in $k$ when $k < d/2$. But alignment summaries are integer points contained in a small volume, and using this fact we have a remarkable result:

**Theorem 2.** For fixed $k$, the number of $k$-best optimality regions is $O(n^{\frac{d(d-1)}{d+1}})$, which matches the best known bound for the $k = 1$ case. Specifically, for general $k$, the number of $k$-best optimality regions is $O((kn)^{\frac{d(d-1)}{d+1}})$ for unordered parametric $k$-best alignment, and $O((k^2n)^{\frac{d(d-1)}{d+1}})$ for ordered parametric $k$-best alignment.

**Remark 1.** Since there might be $\Omega(n^d)$ alignment summaries, Theorem 2 says that, remarkably, the number of $k$-best optimality regions is sublinear in the worst-case number of alignment summaries if $k = o(n^{\frac{1}{1+\epsilon}})$.

In order to leverage Theorem 2 and obtain fast parametric $k$-best alignment, we need to find those very few $k$-best optimality regions, without considering all possible subsets of $k$ summaries. For standard parametric alignment ($k = 1$), the collection of optimality regions can be efficiently represented and computed via an object called the *alignment polytope*. Polytopes are standard geometric objects which generalize polygons to higher dimensions. In [4], polytope construction software was used to efficiently compute alignment polytopes and solve parametric alignment.

For $k > 1$, it was not clear whether $k$-best optimality regions could be represented by a polytope as in the $k = 1$ case. At the heart of our paper is the following affirmative result:

**Theorem 3.** The collection of $k$-best optimality regions can be represented by a polytope called a $k$-set polytope.

Specifically we define *ordered k-set polytopes* and *unordered k-set polytopes*, respectively, for ordered and unordered parametric $k$-best alignment. For $k = 1$ the $k$-set polytopes are precisely the alignment polytope. Our $k$-set polytopes elucidate the structure of $k$-best optimality regions, and allow us to generalize existing
polytope algorithms for parametric alignment to the $k$-best setting. For standard parametric alignment, the oracle-based incremental polytope construction algorithm in [4] repeatedly runs the NW algorithm as a subroutine, with different choices of scoring parameters, in order to find new optimal alignment summaries. Here we generalize the oracle-based incremental polytope construction algorithm to solve parametric $k$-best alignment. In our generalized incremental algorithm, the standard NW algorithm is replaced with a $k$-best version of NW that finds the $k$-best alignment summaries, instead of just the optimal summary. (The running time of the $k$-best NW algorithm is only a factor of $k$ larger than the running time of standard NW). Our main result is:

**Theorem 4.** There is an oracle-based incremental polytope construction algorithm to solve parametric $k$-best alignment. The algorithm solves unordered parametric $k$-best alignment by calling the $k$-best version of the NW algorithm a total of $O((kn)^{(d-1)/d})$ times. For ordered parametric $k$-best alignment, the $k$-best NW algorithm is called $O((k^2 n)^{(2d-1)/d})$ times. Besides NW calls, the rest of the algorithm’s running time is $O((kn)^{(2d-1)/d})$ for unordered parametric $k$-best alignment, and $O((k^2 n)^{(2d-1)/d})$ for ordered.

Furthermore, for $d \leq 3$ and $k = O(n^{1/4})$ the total running time of our algorithm is optimal, i.e. the running time is the same as running the NW algorithm once for each $k$-best optimality region.

The most important feature of our algorithm’s running time is that the dependence on $k$ is polynomial instead of exponential. For small $k$ the running time of parametric $k$-best alignment is comparable to the best known bounds for the running time of standard parametric alignment. In [4] a whole-genome parametric alignment of *Drosophila* is presented, demonstrating how practical parametric alignment can be in practice. Thus we are confident that parametric $k$-best alignment can be performed at the whole-genome scale as well, for not-too-large $k$. The incremental polytope construction software iB4e [12] can be used right out of the box to compute the necessary $k$-set polytopes, once the $k$-best version of the NW algorithm is written.

2 Background on polyhedral geometry

We begin by reviewing basic definitions and facts in polyhedral geometry.

**Definition 1.** *The convex hull* of a set of points $V = \{v_1, \ldots, v_n\} \subset \mathbb{R}^d$ is the set $\text{conv}(V) = \{ \sum c_i v_i \mid \sum c_i = 1, \ c_i \geq 0 \ \forall i \}$. If $\sum c_i = 1$, and all $c_i \geq 0$, we say $\sum c_i v_i$ is a convex combination of $V$.

**Definition 2.** *A polytope* is a convex hull of any finite non-empty $V \subset \mathbb{R}^d$.

The *dimension* of a polytope $P \subset \mathbb{R}^d$ is the dimension of its relative interior as a manifold. To avoid confusion between $d$ and $\dim P$, $d$ is called the *ambient dimension* of $P$.
Definition 3. Given a polytope \( P \subset \mathbb{R}^d \) and a vector \( c \in \mathbb{R}^d \), the face \( F_c \subset P \) is the set \( F_c = \{ x^* \in P \mid c \cdot x^* = \max_{x \in P} c \cdot x \} \). By convention, the empty set is also considered to be a face of \( P \).

Intuitively faces are the bounding extremities of the polytope. Any face of a polytope \( P \) is again a polytope, whose faces are also faces of \( P \). For most choices of \( c \), the face \( F_c \) will be a single point, which is called a vertex of \( P \). The 1-dimensional faces are called edges, and \((\dim P - 1)\)-dimensional faces are called facets.

Definition 4. Given a polytope \( P \subset \mathbb{R}^d \) and a face \( F \subset P \), the normal cone \( N(F) \) is the set of all vectors \( c \) for which \( F_c \supseteq F \).

In other words, the normal cone \( N(F) \) is the set of all vectors \( c \) such that \( F \) weakly maximizes \( c \cdot x \) over \( P \). There is a natural duality between faces and normal cones: for any two faces \( F, G \subset P \) we have \( G \subset F \) if and only if \( N(G) \supset N(F) \). The relative interiors of the normal cones of a polytope partition \( \mathbb{R}^d \), and the collection of normal cones of all faces of \( P \) is called the normal fan of \( P \).

In this paper we will be interested in computing normal cones of vertices of a polytope \( P \). By duality, it suffices to know the facets of \( P \) as we now explain. For simplicity assume \( \dim P = d \). The facets \( F_c \) of \( P \) which contain \( v \) give the set of vectors \( c \) which generate the normal cone \( N(v) \):

\[
N(v) = \mathbb{R}_{\geq 0}\{c \mid v \in F_c, F_c \text{ is a facet of } P \}
\]

For further reading on polytopes, see [23].

Computing convex hulls

Polytopes have been extensively studied in computational geometry, and many algorithms for convex hull construction have been devised [11, 3, 5]. Unfortunately, traditional convex hull algorithms assume a point set \( S \) is explicitly given, for which \( \text{conv}(S) \) is to be computed. In sequence alignment we are presented with a quite different situation. We don’t know the set \( S \), but we seek to compute vertices and facets of \( \text{conv}(S) \), and we have a fast oracle (e.g. the NW algorithm) which will find a vertex of \( \text{conv}(S) \) that maximizes the dot-product with a given vector \( c \). The incremental construction algorithm and software reported in [4, 15] builds convex hulls efficiently in this setting. Briefly put, the incremental construction algorithm repeatedly queries the vertex-finding oracle with different vectors \( c \), adding one vertex at a time to the polytope, until all vertices of the convex hull are guaranteed to be found. As shown in [16], we have

Theorem 5. The incremental construction algorithm builds the convex hull of a point set \( S \), and all faces of \( \text{conv}(S) \), given an oracle \( \text{FindVertex}(c) \) which maximizes given \( c \) over \( S \). The oracle is queried \( O(V + F) \) times, where \( V \) and \( F \) are the number of vertices and facets of \( \text{conv}(S) \). Besides oracle calls, the running time is \( O(\ell_1 + \ldots + \ell_N) \), where \( \ell_j \) is the number of faces of the convex hull after the first \( j \) vertices are added.
Ordered and unordered $k$-set polytopes

It is straightforward to use polytopes as a tool for standard parametric alignment, simply by computing vertices and facets of the alignment polytope, i.e. the convex hull of alignment summaries \[ \mathcal{S} \]. We now present special polytope constructions which are specifically designed for $k$-best alignment. We begin with ordered parametric $k$-best alignment.

We will consider an (implicitly defined, but not explicitly listed) set of $N$ alignment summaries $\mathcal{S} = \{s_i\} \subset \mathbb{Z}^d$, and wish to compute the $k$-best summaries $(s_1, \ldots, s_k)$ in $\mathcal{S}$ with respect to a linear scoring scheme $c \cdot s_1 > c \cdot s_2 > \ldots$. In particular we will be interested in computing all possibilities for the $k$-best summaries as $c$ varies (where the ordering of the summaries is taken into account). We define polytopes $P_k$, which we call ordered $k$-set polytopes, whose vertices correspond to obtainable tuples of $k$-best summaries.

**Definition 5.** Given a set of $N$ alignment summaries $\mathcal{S} = \{s_i\} \subset \mathbb{Z}^d$, let $(N)_k$ denote the set of all $N(N-1) \cdots (N-k+1)$ tuples of $k$ distinct indices $\sigma = (\sigma(1), \ldots, \sigma(k))$, where $\sigma(1), \ldots, \sigma(k) \in \{1, 2, \ldots, N\}$. Notice that the ordering of the indices is taken into account. The ordered $k$-set polytope $P_k$ for $\mathcal{S}$ is the convex hull

$$P_k = \text{conv} \{ \sum_{i=1}^k (k+1-i)s_{\sigma(i)} | \sigma \in (N)_k \}.$$

| Example of definition of ordered 3-set polytope $P_3$, when $\mathcal{S}$ is a set of four points $s_1, s_2, s_3, s_4$. In this case $P_3$ is the convex hull of $(4)_3 = 4 \cdot 3 \cdot 2 = 24$ points. The 24 points are listed above. |
|---|
| $3s_1 + 2s_2 + s_3$ | $3s_2 + 2s_1 + s_3$ | $3s_3 + 2s_1 + s_2$ | $3s_4 + 2s_1 + s_2$ |
| $3s_1 + 2s_2 + s_4$ | $3s_2 + 2s_1 + s_4$ | $3s_3 + 2s_1 + s_4$ | $3s_4 + 2s_1 + s_3$ |
| $3s_1 + 2s_3 + s_2$ | $3s_2 + 2s_3 + s_1$ | $3s_3 + 2s_2 + s_1$ | $3s_4 + 2s_2 + s_1$ |
| $3s_1 + 2s_3 + s_4$ | $3s_2 + 2s_3 + s_4$ | $3s_3 + 2s_2 + s_4$ | $3s_4 + 2s_2 + s_3$ |
| $3s_1 + 2s_4 + s_2$ | $3s_2 + 2s_4 + s_1$ | $3s_3 + 2s_4 + s_1$ | $3s_4 + 2s_3 + s_1$ |
| $3s_1 + 2s_4 + s_3$ | $3s_2 + 2s_4 + s_3$ | $3s_3 + 2s_4 + s_2$ | $3s_4 + 2s_3 + s_2$ |

The following theorem shows that the normal fan of $P_k$ gives precisely the ordered $k$-best optimality regions for the alignment summaries $\mathcal{S}$. Thus computing vertices and facets of $P_k$ completely solves ordered parametric $k$-best alignment.

**Theorem 6.** The normal cone of a point $\sum_{i=1}^k (k+1-i)s_{\sigma(i)} \in P_k$ is the set of all $c$ satisfying $c \cdot s_{\sigma(1)} \geq c \cdot s_{\sigma(2)} \geq \ldots \geq c \cdot s_{\sigma(k)}$, and $c \cdot s_{\sigma(j)} \geq c \cdot s_j$ for all $j \notin \sigma$.

**Proof.** See [14].

We now give analogous results for unordered parametric $k$-best alignment.
Definition 6. Given a set of \( N \) points \( S = \{ s_i \} \subset \mathbb{R}^d \), let \( {S \choose k} \) denote the set of all \( \binom{N}{k} \) subsets of \( S \) of size \( k \). The unordered \( k \)-set polytope \( Q_k \) for \( S \) is

\[
Q_k = \text{conv}\left\{ \sum_{s \in A} s \mid A \in {S \choose k} \right\}.
\]

Unordered \( k \)-set polytopes have been previously studied \cite{6,1}. We can modify Theorem 6 to show that the normal fan of \( Q_k \) gives the (unordered) \( k \)-best optimality regions for alignment summaries:

Theorem 7. The normal cone of a point \( \sum_{s \in A} s \in Q_k \) is the set of vectors \( c \) which satisfy \( c \cdot s \geq c \cdot s' \) for all \( s \in A \) and \( s' \notin A \).

Example 1. Suppose \( S \) is the set of four vertices of a square. Figure 1 shows some unordered \( k \)-set polytopes \( Q_k \) and ordered \( k \)-set polytopes \( P_k \) for \( S \).

![Figure 1: Examples of unordered k-set polytopes Q_k and ordered k-set polytopes P_k, for S = \{a, b, c, d\} = vertices of a square. Points are labeled by the ordered/unordered k-set they represent. Notice for example the point 2b + d is in the interior of P_2; this means that it is impossible for a linear scoring scheme to make (b, d) the ordered top-2 points.](image)

In parametric \( k \)-best alignment, the set of points \( S \) are alignment summaries, which are integer points. In this case we can obtain remarkable bounds on the complexity of \( k \)-set polytopes.

Theorem 8. Suppose \( S \subset \mathbb{Z}^d \) is a set of \( N \) integer points, and \( k < N \). Let \( V \) be the volume of \( \text{conv}(S) \), and assume \( V > 0 \). If \( V \) is any subset of vertices of the unordered \( k \)-set polytope \( Q_k \) for \( S \), then the total number of faces of \( \text{conv}(V) \) is \( O\left((k^dV)^{(d-1)/(d+1)}\right) \). Similarly, if \( W \) is any subset of the vertices of ordered \( k \)-set polytope for \( S \), then the total number of faces of \( \text{conv}(W) \) is \( O\left((k^{2d}V)^{(d-1)/(d+1)}\right) \).

Proof. See Appendix.

This concludes our treatment of \( k \)-set polytopes. The rest of the paper gives applications to parametric \( k \)-best alignment, along with details on computation and implementation.
4 Parametric $k$-best alignment

We now aggregate the results of the previous sections to efficiently solve parametric $k$-best alignment. For given sequences $\sigma_1, \sigma_2$ of length $O(n)$, let $S = \{s_1, \ldots, s_N\}$ be the set of all alignment summaries. Each entry in an alignment summary counts a feature in the alignment, such as the number of occurrences of a type of mismatch. Thus $\text{conv}(S)$ has volume $O(n^d)$.

We first explain how to solve ordered parametric $k$-best alignment. For $k \leq N$ let $P_k$ be the ordered $k$-set polytope for $S$. Given values for alignment scoring parameters $c$, the NW algorithm finds the optimal alignment summary by dynamic programming, keeping track of the optimal summary at each node in the alignment graph \cite{20}. Similarly, the NW algorithm can compute the top $k$ distinct alignment summaries using the same type of dynamic programming recursion, keeping track of the top $k$ distinct summaries at each node in the alignment graph. Thus we can define an oracle which, given $c$, will find the vertex $s^* \in P_k$ which maximizes $c \cdot s$ over $P_k$:

- Call the NW algorithm with scoring parameters $c$, to compute the top $k$ distinct alignment summaries $s_1, \ldots, s_k$ such that $c \cdot s_1 > \ldots > c \cdot s_k > \ldots$.
- Return $s^* := k s_1 + (k-1)s_2 + \cdots + s_k$.

Then Theorem \ref{thm:polytope-construction} says that the incremental polytope construction method will compute $P_k$ and its normal fan, calling the above oracle $O(V + F)$ times where $V$ and $F$ are the number of vertices and facets of $P_k$. Since $\text{conv}(S)$ has volume $O(n^d)$, Theorem \ref{thm:polytope-construction} says that $V$ and $F$ are both $O((k^2 n^{(d-1)/d+1}))$. Theorems \ref{thm:polytope-construction} and \ref{thm:polytope-construction} also tell us that besides oracle calls, the incremental polytope construction method takes no more than $O(V(k^2 n^{(d-1)/(d+1)}))$ time, which is $O((k^2 n^{2(d-1)/(d+1)})$. Putting it all together, we have

**Theorem 9.** Given sequences of length $O(n)$, let $V, F$ be the number of vertices and facets of the ordered $k$-set polytope for alignment summaries. The incremental polytope construction method will solve ordered parametric $k$-best alignment in $O((V + F)W(n, k) + (k^2 n^{2d(d-1)/(d+1)}))$ time, where $W(n, k)$ is the time required to run $k$-best NW once on sequences of length $O(n)$.

Typically $W(n, k) = \Theta(kn^2)$. Table \ref{table:complexity} gives complexity bounds in this case for specific small values of $d$.

**Remark 2.** The upper bound theorem for polytopes \cite{23} says that the number of faces of a $d$-dimensional polytope is linear in the number of vertices if $d \leq 3$. So if $d \leq 3$, and $W(n, k) = \Theta(kn^2)$, the running time of our algorithm is $O(V kn^2 + V^2)$, and since $V = O(k^3 n^{3/2})$, the running time is thus $O(V \cdot W(n, k))$ when $k = O(n^{1/4})$. This is the same running time as the time required to run the NW algorithm once for each top $k$ ranking. Thus for $d \leq 3$ and $k = O(n^{1/4})$, our algorithm is an optimal oracle-based method.

**Remark 3.** When $d = 2$, at most one facet (edge) of the ordered $k$-set polytope is deleted when a new vertex is added (otherwise at least one vertex $v$ would also be
deleted, contradicting that $v$ is a vertex). Thus for $d = 2$ and $W(n, k) = \Theta(kn^2)$ the running time is always the optimal $O(V \cdot W(n, k)) = O(k^{7/3} n^{8/3})$ for any $k$.

**Remark 4.** In practice, we have observed that relatively few faces are created or destroyed when each new vertex of $P_k$ is found. If an amortized $O(1)$ faces are created or destroyed when each new vertex of $P_k$ is found, then the incremental polytope construction method solves ordered parametric $k$-best alignment in $O((V + F) \cdot W(n, k))$ time, which is $O((k^2 n)^d/d+1) \cdot W(n, k))$. It is an important open question to determine the worst-case number of faces that can be created or destroyed during the incremental construction of $P_k$.

We now explain how to solve unordered parametric $k$-best alignment. The solution is analogous to ordered $k$-best alignment. Let $Q_k$ be the unordered $k$-set polytope for $S$. We define an oracle which, given $c$, will find the vertex $s^* \in Q_k$ which maximizes $c \cdot s$ over $Q_k$:

- Call the NW algorithm with alignment scoring parameters $c$, to compute the top $k$ distinct alignment summaries $s_1, \ldots, s_k$ such that $c \cdot s_1 > \ldots > c \cdot s_k > \ldots$.
- Return $s^* := s_1 + s_2 + \cdots + s_k$.

Then, endowed with the above vertex-finding oracle, we have

**Theorem 10.** For sequences of length $O(n)$, let $V, F$ be the number of vertices and facets of the unordered $k$-set polytope for alignment summaries. The incremental polytope construction method will solve unordered parametric $k$-best alignment in $O((V + F) W(n, k) + (kn)^{d(d-1)/2})$ time.

Table 2 gives specific bounds for small values of $d$, assuming $W(n, k) = \Theta(kn^2)$.

| $d$   | Output size (unordered) | Running time (unordered) | Output size (ordered) | Running time (ordered) |
|-------|-------------------------|--------------------------|-----------------------|------------------------|
| 2     | $O(k^{2/3} n^{2/3})$   | $O(k^{5/3} n^{8/3})$    | $O(k^{4/3} n^{2/3})$ | $O(k^{7/3} n^{8/3})$  |
| 3     | $O(k^{3/2} n^{3/2})$   | $O(k^{5/2} n^{7/2} + k^3 n^3)$ | $O(k^3 n^{3/2})$     | $O(k^4 n^{7/2} + k^6 n^3)$ |
| 4     | $O(k^{12/5} n^{12/5})$ | $O(k^{24/5} n^{24/5})$  | $O(k^{24/5} n^{12/5})$ | $O(k^{48/5} n^{24/5})$ |

Table 2: Running time and output complexity of ordered/unordered parametric $k$-best alignment for small dimensions, assuming the $k$-best version of the NW algorithm runs in $\Theta(kn^2)$ time.

**Remark 5.** Analogous to Remark 4, the running time would be $O((V + F) \cdot W(n, k))$ if we could prove that an amortized $O(1)$ faces are created or destroyed when each new vertex of $Q_k$ is found.

**Remark 6.** Analogous to Remark 4, if $d \leq 3$ the running time is the optimal $O(V \cdot W(n, k))$ when $k = O(n)$ and $W(n, k) = \Theta(kn^2)$. 

9
The software iB4e reported in [15] can be used right out of the box to solve ordered and unordered parametric k-best alignment this way, once the oracles for finding vertices of $P_k$ and $Q_k$ are written. A beta version of iB4e was used in [4] to perform high throughput parametric alignment, and greatly outperformed the “polytope semiring” method reported in [19].

5 Discussion

Although parametric alignment is a major improvement upon standard sequence alignment, parametric alignment ignores nearly optimal alignments. Here we have extended parametric alignment to the $k$-best setting, determining how the top $k$ alignment summaries vary with scoring parameters. This allows for much more realistic parametric analysis of biological sequences.

Parametric alignment has enjoyed remarkably good complexity results, enabling whole-genome parametric analysis of Drosophila genomes [4]. By extending the good complexity results to parametric $k$-best alignment, we believe parametric $k$-best alignment can be performed at the whole-genome scale as well. As in [4], such genome-scale parametric analysis will require standard preprocessing techniques that break up pairwise genomes into smaller reliably homologous subsequences.

In some applications, estimates of scoring parameters might be known along with confidence intervals on the estimates. In this case we can restrict attention to optimality regions which intersect the confidence region for parameters. It is possible to augment the vertex-finding oracle in iB4e so that only optimal alignments whose optimality regions intersect a prescribed cone $C$ are found; other optimal alignments are completely avoided. Details can be found in [16]. For example optimality regions could be restricted to a cone over a bounding box, as in [13]. Restricting the parameter space has the additional benefit of speeding up parametric $k$-best alignment, by reducing the number of optimality regions.

The dimension $d$ of alignment summaries is the most prohibitive factor in the complexity of both parametric alignment and $k$-best alignment. But the curse of dimension is not nearly as bad as was speculated in [4]. While some polytopes with $V$ vertices might have $\Theta(V^{\lfloor d/2 \rfloor})$ faces, we have shown that $k$-set polytopes are special, and that the remarkable bounds on their number of vertices also applies to faces of all dimensions. Thus parametric alignment and $k$-best alignment are much more tractable than previously thought. This agrees with empirical observations, e.g. in [4] parametric alignment was demonstrated to be computationally practical for $d \leq 5$ at the whole genome level. Our complexity results indicate that parametric $k$-best alignment will be similarly tractable. Based on computational experience with parametric alignment, we believe parametric $k$-best alignment will even be tractable for $d = 6, 7$ when sequences are short.

It is important to note that restricting alignment summaries to have dimension $\leq 7$ prohibits the most general models of alignment scoring parameters. For protein sequences, all but the most basic scoring matrices will yield $d > 7$. Thus parametric alignment is not well-suited for protein sequence analysis. Fortunately, for DNA sequences, popular scoring models such as those based on Jukes–Cantor, Kimura-2,
and Kimura-3 scoring matrices will result in $d \leq 6$.

Parametric alignment belongs to a more general class of algorithms called parametric inference algorithms for graph-based models \[19\]. We note that the framework we have laid out here, extending parametric alignment to the $k$-best setting, can be adapted to perform parametric $k$-best inference in other graph-based models as well. The remarkable complexity results we have proved can be extended to the parametric $k$-best inference setting as well. Similarly the software \textsc{iB4e} can be used to perform efficient parametric $k$-best inference, provided an oracle which performs $k$-best inference given scoring parameters. Two important graph-based models in biology which can benefit from parametric $k$-best inference are hidden Markov models over discrete state spaces, and tree-models for single nucleotide evolution. The vertex-finding oracles provided to \textsc{iB4e} for these graph-based models would be the $k$-best Viterbi algorithm and $k$-best Felsenstein pruning algorithm respectively.

\section*{Acknowledgements}

R. Yoshida is supported by NIH R01 grant 1R01GM086888-01.
References

[1] A Andrzejak and K Fukuda. Optimization over k-set polytopes and efficient k-set enumeration. WADS, pages 1–12, 1999.

[2] I Bárány and D G Larman. The convex hull of the integer points in a large ball. Math. Annalen, 312:167–181, 1998.

[3] B C Barber, D P Dobkin, and H Huhdanpaa. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software, 22(4):469–483, 1996.

[4] C Dewey, P Huggins, L Pachter, B Sturmfels, and K Woods. Parametric alignment of drosophil genomes. PLoS Computational Biology, 2(6):e73, 2006.

[5] H Edelsbrunner. Algorithms in Combinatorial Geometry, volume 10. Monographs in Theoretical Computer Science. An EATCS Series, Vienna, 1987.

[6] H Edelsbrunner, P Valtr, and E Welzl. Cutting dense point sets in half. Discrete Comput. Geom., 17:243–255, 1997.

[7] D Fernández-Baca, T Seppäläinen, and G Slutzki. Bounds for parametric sequence comparison. Discrete Applied Mathematics, 118:181–198, 2002.

[8] D Fernández-Baca, T Seppäläinen, and G Slutzki. Parametric multiple sequence alignment and phylogeny construction. Journal of Discrete Algorithms, 2(2):271–287, 2004.

[9] D Fernández-Baca and B Venkatachalam. Sensitivity analysis in combinatorial optimization, 2007. In Handbook of Approximation Algorithms and Metaheuristics. T. Gonzalez, ed., Chapman and Hall/CRC Press Computer and Information Science Series.

[10] D Fernández-Baca and S B Venkatachalam. Parametric sequence alignment, 2005. Handbook of computational molecular biology Computer and Information Science Series (Volume 2): Chapman and Hall/CRC Press. 271287.Computer and Information Science Series.

[11] K Fukuda, T H Liebling, and C Lütlof. Extended convex hull. Comput. Geom., 20:13–23, 2001.

[12] D Gusfield, K Balasubramanian, and D Naor. Parametric optimization of sequence alignment. Algorithmica, 12:312–326, 1994.

[13] D Gusfield and P Stelling. Parametric and inverseparametric sequence alignment with xparal. Methods Enzymol, 266:481–494, 1996.

[14] P Huggins. Ordered and unordered k-set polytopes. Progress.
[15] P Huggins. ib4e: A software framework for parametrizing specialized lp problems, 2006. A Iglesias, N Takayama (Eds.): Mathematical Software - ICMS. Second International Congress on Mathematical Software, Castro Urdiales, Spain, September 1-3, Proceedings. Lecture Notes in Computer Science 4151 Springer.

[16] P Huggins. Polytopes in computational biology, 2008. PhD Thesis.

[17] D. Naor and D. Brutlag. On suboptimal alignments of biological sequences. In A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, editors, Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching, number 684, pages 179–196, Padova, Italy, 1993. Springer-Verlag, Berlin.

[18] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal Molecular Biology, 48:443–445, 1970.

[19] L Pachter and B Sturmfels. Parametric inference for biological sequence analysis. Proc Natl Acad Sci U S A, 101(46):16138–43, 2004.

[20] L Pachter and B Sturmfels. Algebraic Statistics for Computational Biology. Cambridge University Press, 2005.

[21] Martin Vingron. Near-optimal sequence alignment. Current Opinion in Structural Biology, 6(3):346–352, June 1996.

[22] M Waterman, M Eggert, and E Lander. Parametric sequence comparisons. Proc Natl Acad Sci USA, 89:6090–6093, 1992.

[23] G Ziegler. Lectures on Polytopes. Springer-Verlag, Vienna, 1995.
A Appendix

Ordered and unordered $k$-set polytopes have rich structure which has still not been fully explored. Here we recall a result from [14] which shows that $k$-set polytopes can be intractable for general sets of points:

**Proposition 1.** Given a set of $N$ points $S \subset \mathbb{R}^d$, let $f_k$ be the number of $k$-dimensional faces of $\text{conv}(S)$. Then the ordered $k$-set polytope for $V$ has at least $(k+1)!f_k$ vertices. Thus $k$-set polytopes have $\Theta(N^k)$ vertices in the worst case, if $k < d/2$.

However, alignment summaries are integer points. Suppose $S \subset \mathbb{Z}^d$ is a finite set of integer points. Then as recalled in [2], we have:

**Theorem 11** (Andrews–Barany). Let $V$ be the volume of $\text{conv}(S)$. If $V > 0$, then the number of $k$-dimensional faces of $\text{conv}(S)$ is $O(V^{(d-1)/(d+1)})$ for every $k$.

No such result is possible when $S$ is an arbitrary set of real-valued points.

A.1 Proof of Theorem [8]

The proof requires a lemma, proved in [14]:

**Lemma 1.** Suppose $S \subset \mathbb{R}^d$ is a set of $N \geq 1$ points. If $1 \leq k < N$, then the $k$-set polytopes $Q_k$ and $P_k$ for $S$ have the same dimension as $\text{conv}(S)$.

**Proof of Theorem [8]** By definition of Minkowski sum, every point in the $k$-set polytope $Q_k$ is also a point in the $k$-fold Minkowski sum $\text{conv}(S)^{\odot k}$, which equals the $k$-fold dilation $k \cdot \text{conv}(S) = \{kx \mid x \in \text{conv}(S)\}$. Thus the volume of $Q_k$ is no more than the volume of $k \cdot \text{conv}(S)$, which is $k^d V$. By the lemma, the volume of $Q_k$ is positive, so we can apply Theorem [11] and obtain that the number of faces of $Q_k$ is $O((k^d V)^{(d-1)/(d+1)})$.

Now, for the ordered $k$-set polytope $P_k$, we recall that $P_k = Q_1 \odot \ldots \odot Q_k$. Since each $Q_j$ is a subset $j \cdot \text{conv}(S)$, we have that $P_k \subset k^2 \cdot \text{conv}(S)$. Now $P_k$ is a lattice polytope of positive volume $\leq k^{2d} V$. So Theorem [8] tells us that the number of faces of $P_k$ is $O((k^{2d}V)^{(d-1)/(d+1)})$. 

\hfill $\square$