Combinatorics from biology: inference functions and sequence alignment

Sergi Elizalde

`elizalde@msri.org`

MSRI / Dartmouth College

Cambridge University Press is pleased to announce...

Algebraic Statistics for Computational Biology

Lior Pachter, *University of California, Berkeley*
Bernd Sturmfels, *University of California, Berkeley*

The quantitative analysis of biological sequence data is based on methods from statistics coupled with efficient algorithms from computer science. Algebra provides a framework for unifying many of the seemingly disparate techniques used by computational biologists. This book offers an introduction to this mathematical framework and describes tools from computational algebra for designing new algorithms for exact, accurate results. These algorithms can be applied to biological problems such as aligning genomes, finding genes, and constructing phylogenies.

The first part of this book consists of four chapters on the themes of Statistics, Computation, Algebra, and Biology offering speedy, self-contained introductions to the emerging field of algebraic statistics and its applications to genomics. In the second part, the four themes are combined and developed to tackle real problems in computational genomics. As the first book in this exciting and dynamic area, it will be welcomed as a text for self-study or for advanced undergraduate and beginning graduate courses.

CONTENTS:

Preface; Part I. Introduction to the Four Themes: 1. Statistics L. Pachter and B. Sturmfels; 2. Computation L. Pachter and B. Sturmfels; 3. Algebra L. Pachter and B. Sturmfels; 4. Biology L. Pachter and B. Sturmfels; Part II. Studies on the Four Themes: 5. Parametric inference R. Mihaescu; 6. Polytope propagation on graphs M. Joswig; 7 Parametric sequence alignment C. Dewey and K. Woods; 8. Bounds for optimal sequence alignment S. Elizalde; 9. Inference functions S. Elizalde; 10. Geometry of Markov chains E. Kuo; 11. Equations defining Hidden Markov Models N. Bray and J. Morton; 12. The EM algorithm for Hidden Markov Models I.B. Hallgrímsdóttir, A. Milowski and J. Yu; 13. Homology mapping with Markov random fields A. Caspi; 14. Mutagenetic tree models N. Beerenwinkel and M. Drton; 15. Catalog of small trees M. Casanellas, L. Garcia, and S. Sullivant; 16. The strand symmetric model M. Casanellas and S. Sullivant; 17. Extending statistical models from trees to splits graphs D. Bryant; 18. Small trees and generalized neighbor-joining M. Contois and D. Levy; 19. Tree construction using Singular Value Decomposition N. Eriksson; 20. Applications of interval methods to phylogenetics R. Sainudiin and R. Yoshida; 21. Analysis of point mutations in vertebrate genomes J. Al-Aidroos and S. Snir; 22. Ultra-conserved elements in vertebrate genomes M. Drton, N. Eriksson and G. Leung; Index

July 2005 / c. 400 pages / 100 line diagrams / 5 tables / 3 color plates
Hardback / 0-521-85700-7 / List: $60.00 /* ***Discount: $48.00****

- In biology, an important source of data is the genome, which is a long sequence of letters `A`, `C`, `G` and `T`.

- In biology, an important source of data is the genome, which is a long sequence of letters `A`, `C`, `G` and `T`.
- In computational biology, one tries to analyze this data and obtain information from it:
    - evolution of species (phylogenetic trees),
    - structure of the genome,
    - ...

Introduction

- In biology, an important source of data is the genome, which is a long sequence of letters $\mathtt{A}, \mathtt{C}, \mathtt{G}$ and $\mathtt{T}$.
- In computational biology, one tries to analyze this data and obtain information from it:
  - evolution of species (phylogenetic trees),
  - structure of the genome,
  - ...
- Statistical models are used to analyze these sequences and draw conclusions from them.

Example 1: sequence alignment

- It is used to determine the similarity between biological sequences.
- It gives information about how a sequence evolved into another by a sequence of insertions, deletions, and mutations.
- Given two sequences, the problem is to find an optimal alignment.

Example 1: sequence alignment

- It is used to determine the similarity between biological sequences.
- It gives information about how a sequence evolved into another by a sequence of insertions, deletions, and mutations.
- Given two sequences, the problem is to find an optimal alignment.

Example:

Consider the sequences $S = \mathtt{AGGAATTCG}$ and $T = \mathtt{GGACGGTAT}$.

Example 1: sequence alignment

- It is used to determine the similarity between biological sequences.
- It gives information about how a sequence evolved into another by a sequence of insertions, deletions, and mutations.
- Given two sequences, the problem is to find an optimal alignment.

Example:

Consider the sequences $S = \mathtt{AGGAATTCG}$ and $T = \mathtt{GGACGGTAT}$.

Which of the following two alignments is better?

$S_1 = \mathtt{AGGA{-}ATTCG}$	$S_2 = \mathtt{{-}{-}A{-}GGAATTCG}$
$T_1 = \mathtt{{-}GGACGGTAT}$	$T_2 = \mathtt{GGACGGTAT{-}{-}{-}}$

Example 1: sequence alignment

$S_1 =$ AGGA-ATTCG

$T_1 =$ -GGACGGTAT

$S_2 =$ --A-GGAATTCG

$T_2 =$ GGACGGTAT---

Example 1: sequence alignment

$S_1 =$ `AGGA-ATTCG`	$S_2 =$ `--A-GGAATTCG`
$T_1 =$ `-GGACGGTAT`	$T_2 =$ `GGACGGTAT---`

4	$\longleftarrow$	$z = \#$matches	$\longrightarrow$	5
4	$\longleftarrow$	$x = \#$mismatches	$\longrightarrow$	1
1	$\longleftarrow$	$y = \frac{1}{2}\#$spaces$= \#$insertions	$\longrightarrow$	3

Example 1: sequence alignment

$$S_1 = \texttt{AGGA-ATTCG} \qquad\qquad S_2 = \texttt{--A-GGAATTCG}$$

$$T_1 = \texttt{-GGACGGTAT} \qquad\qquad T_2 = \texttt{GGACGGTAT---}$$

4	$\longleftarrow$	$z = \#$matches	$\longrightarrow$	5
4	$\longleftarrow$	$x = \#$mismatches	$\longrightarrow$	1
1	$\longleftarrow$	$y = \frac{1}{2}\#$spaces$= \#$insertions	$\longrightarrow$	3

One possible model: matches are rewarded by 1,
mismatches are penalized by α,
insertions are penalized by β.

$4 - 4\alpha - \beta \quad \longleftarrow \quad$ score:$= z - x\alpha - y\beta \quad \longrightarrow \quad 5 - \alpha - 3\beta$

- An *optimal alignment* is one with highest score.

  Optimality depends on the values of the parameters α and β.

  Example:

  $$4 - 4\alpha - \beta > 5 - \alpha - 3\beta \quad \text{if} \quad \alpha = 0.3, \;\; \beta = 1, \quad \text{but}$$
  $$4 - 4\alpha - \beta < 5 - \alpha - 3\beta \quad \text{if} \quad \alpha = 0.4, \;\; \beta = 0.7.$$

- An *optimal alignment* is one with highest score.

  Optimality depends on the values of the parameters α and β.

  Example:

  $$4 - 4\alpha - \beta > 5 - \alpha - 3\beta \quad \text{if} \quad \alpha = 0.3, \;\; \beta = 1, \quad \text{but}$$
  $$4 - 4\alpha - \beta < 5 - \alpha - 3\beta \quad \text{if} \quad \alpha = 0.4, \;\; \beta = 0.7.$$

- For fixed parameters, we get a function that maps each pair of sequences to their optimal alignment.

  Such a map is called an *inference function*.

Example 1: sequence alignment

- An *optimal alignment* is one with highest score.

  Optimality depends on the values of the parameters α and β.

  Example:

  $$4 - 4\alpha - \beta > 5 - \alpha - 3\beta \quad \text{if} \quad \alpha = 0.3, \;\; \beta = 1, \quad \text{but}$$
  $$4 - 4\alpha - \beta < 5 - \alpha - 3\beta \quad \text{if} \quad \alpha = 0.4, \;\; \beta = 0.7.$$

- For fixed parameters, we get a function that maps each pair of sequences to their optimal alignment.

  Such a map is called an *inference function*.

- The number of maps from pairs of sequences of length n to alignments is doubly exponential in n.

  However, most of them are not inference functions for any choice of parameters.

Example 1: sequence alignment

- An *optimal alignment* is one with highest score.

  Optimality depends on the values of the parameters α and β.

  Example:

  $4 - 4\alpha - \beta > 5 - \alpha - 3\beta$ if $\alpha = 0.3, \ \beta = 1,$ but
  $4 - 4\alpha - \beta < 5 - \alpha - 3\beta$ if $\alpha = 0.4, \ \beta = 0.7.$

- For fixed parameters, we get a function that maps each pair of sequences to their optimal alignment.

  Such a map is called an *inference function*.

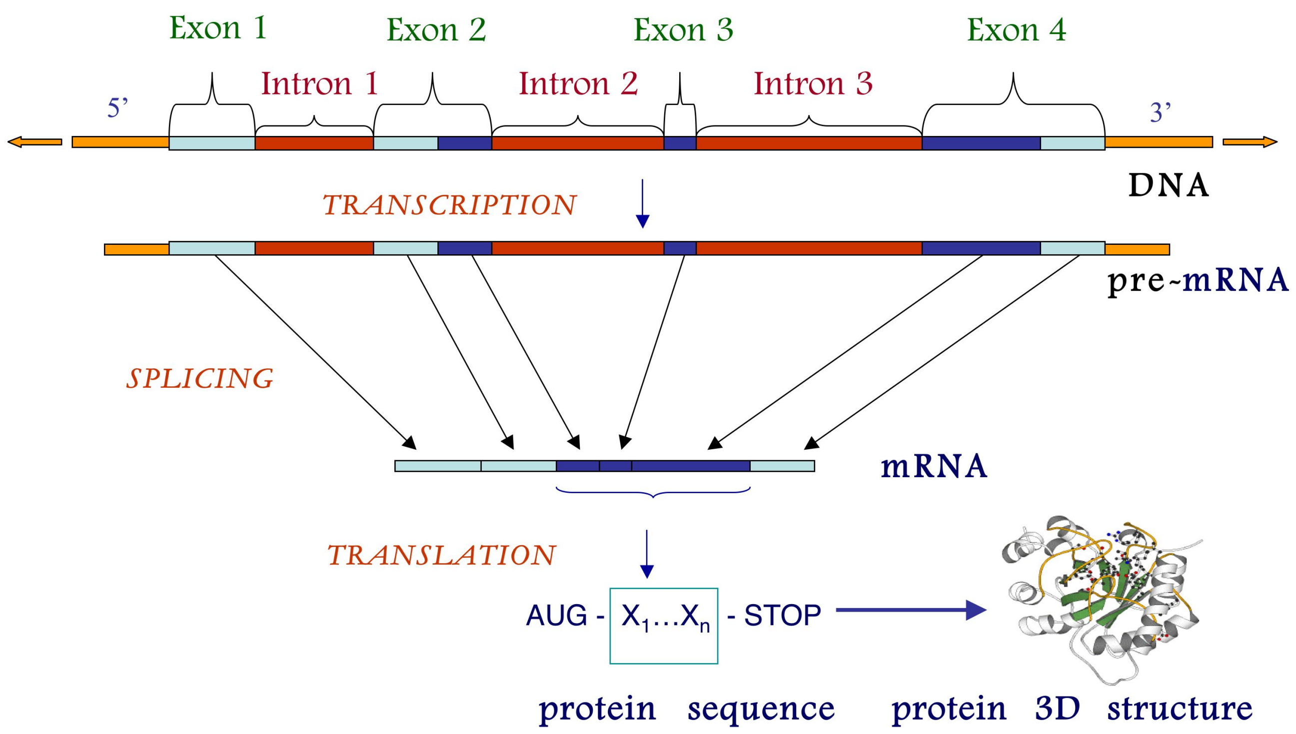
- The number of maps from pairs of sequences of length n to alignments is doubly exponential in n.

  However, most of them are not inference functions for any choice of parameters.

Problem: How many of them are inference functions?

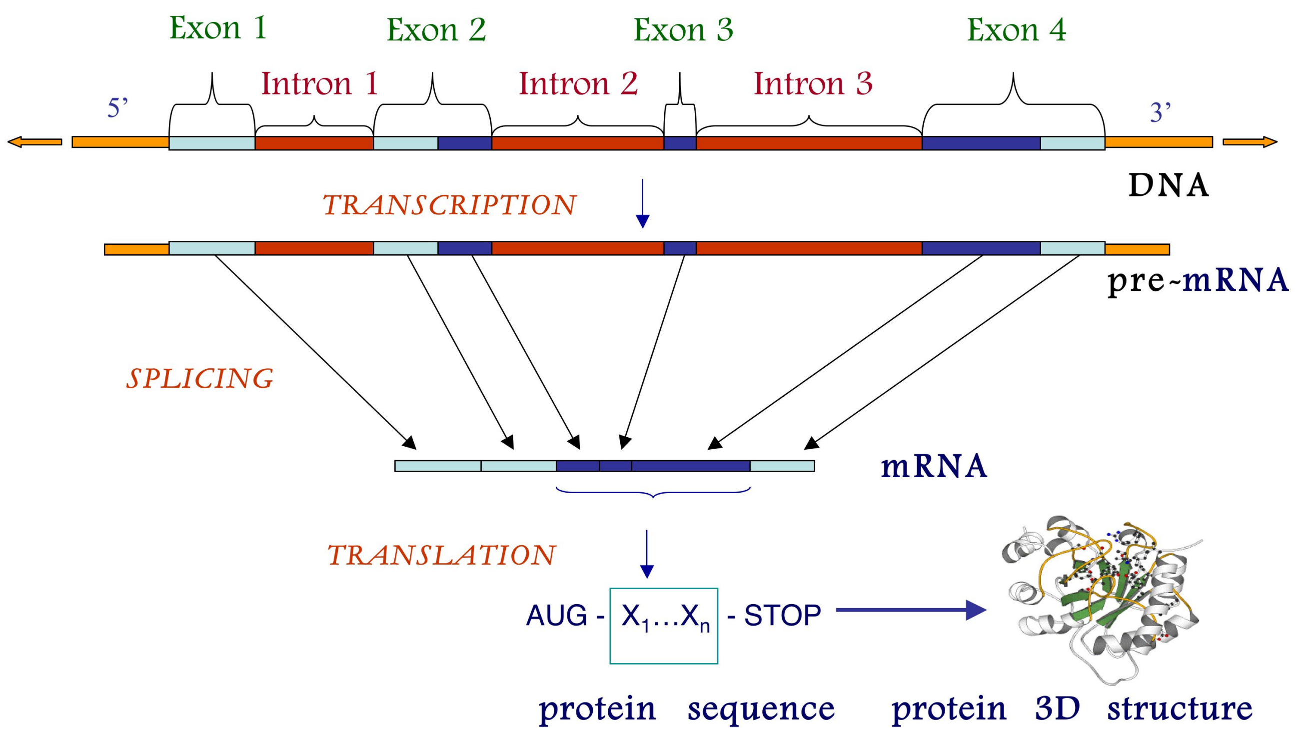
Example 2: gene finding functions

A problem in biology is to distinguish introns from exons in the genome.



Example 2: gene finding functions

A problem in biology is to distinguish introns from exons in the genome.



Exons are translated intro proteins, but introns are spliced out before the translation.

Example 2: gene finding functions

- A *gene finding function* is a map

$$\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^n \longrightarrow \{\mathrm{in}, \mathrm{ex}\}^n$$

that, for each sequence of DNA, outputs its “most likely” division into introns and exons.

Example 2: gene finding functions

- A *gene finding function* is a map

$$\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^n \longrightarrow \{\mathrm{in}, \mathrm{ex}\}^n$$

  that, for each sequence of DNA, outputs its “most likely” division into introns and exons.

- They are inference functions of a certain model.

Example 2: gene finding functions

- A *gene finding function* is a map

$$\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^n \longrightarrow \{\mathrm{in}, \mathrm{ex}\}^n$$

  that, for each sequence of DNA, outputs its “most likely” division into introns and exons.

- They are inference functions of a certain model.

- Even though the total number of maps $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^n \longrightarrow \{\mathrm{in}, \mathrm{ex}\}^n$ is 2^{n4^n}, very few of them are gene finding functions
  (i.e., they don't maximize any reasonable scoring function).

Example 2: gene finding functions

- A *gene finding function* is a map

$$\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^n \longrightarrow \{\mathrm{in},\mathrm{ex}\}^n$$

  that, for each sequence of DNA, outputs its “most likely” division into introns and exons.

- They are inference functions of a certain model.

- Even though the total number of maps $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^n \longrightarrow \{\mathrm{in},\mathrm{ex}\}^n$ is 2^{n4^n}, very few of them are gene finding functions
  (i.e., they don't maximize any reasonable scoring function).

Problem: How many of them are gene finding functions?

Algebraic statistical models

They are a general setting used to analyze biological data.

They are a general setting used to analyze biological data.

An *algebraic statistical model* arises as the image of a polynomial map

$$\mathbf{f}: \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R}^m \\ \underbrace{\theta = (\theta_1, \ldots, \theta_d)}_{\text{model parameters}} & \mapsto & \underbrace{(f_1(\theta), f_2(\theta), \ldots, f_m(\theta))}_{\text{probability distribution}} \end{array}$$

$$f_i(\theta) \geq 0$$

$$f_1(\theta) + f_2(\theta) + \cdots + f_m(\theta) = 1$$

Algebraic statistical models

They are a general setting used to analyze biological data.

An *algebraic statistical model* arises as the image of a polynomial map

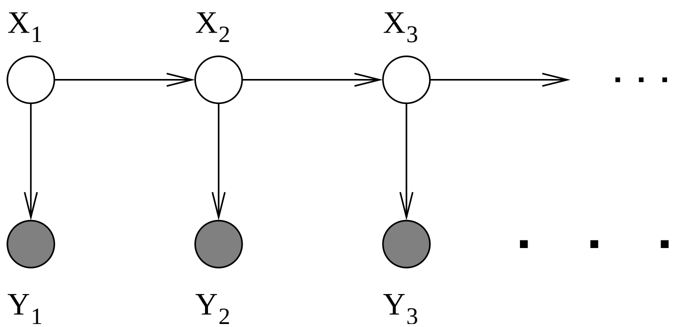
$$\begin{array}{llcl} \mathbf{f}: & \mathbb{R}^d & \longrightarrow & \mathbb{R}^m \\ & \underbrace{\theta = (\theta_1, \ldots, \theta_d)}_{\text{model parameters}} & \mapsto & \underbrace{(f_1(\theta), f_2(\theta), \ldots, f_m(\theta))}_{\text{probability distribution}} \end{array}$$

$$f_i(\theta) \geq 0$$

$$f_1(\theta) + f_2(\theta) + \cdots + f_m(\theta) = 1$$

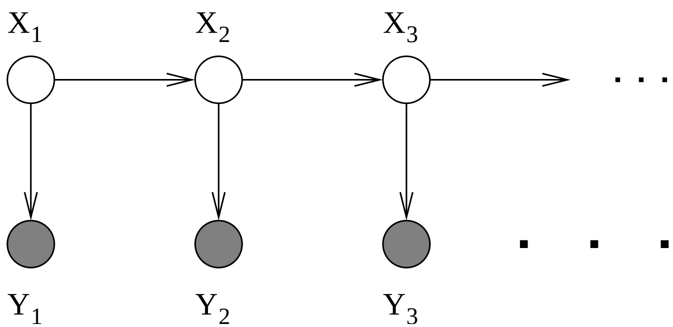
Each coordinate f_i is a polynomial in the d parameters, representing the probability of a certain outcome.

Example: hidden Markov model



} *hidden variables* $X_i \in \Sigma$

} *observed variables* $Y_i \in \Sigma'$



hidden variables $X_i \in \Sigma$

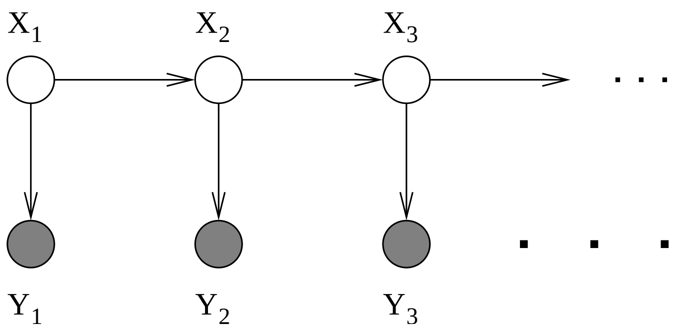
observed variables $Y_i \in \Sigma'$

Transition probabilities: for $i, j \in \Sigma$, $\ell \in \Sigma'$,

$$\theta_{ij} = \mathrm{Prob}(X_{k+1} = j \mid X_k = i)$$

$$\theta_{i\ell} = \mathrm{Prob}(Y_k = \ell \mid X_k = i)$$

Example: hidden Markov model



$\Big\}$ *hidden variables* $X_i \in \Sigma$

$\Big\}$ *observed variables* $Y_i \in \Sigma'$

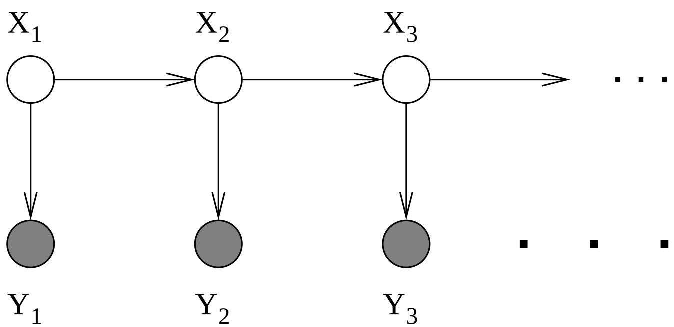
Transition probabilities: for $i, j \in \Sigma$, $\ell \in \Sigma'$,

$$\theta_{ij} \;=\; \mathrm{Prob}(X_{k+1} = j \mid X_k = i)$$

$$\theta_{i\ell} \;=\; \mathrm{Prob}(\,Y_k = \ell \mid X_k = i)$$

- The only information that we see are the values of $Y_1, Y_2, \ldots$. We want to guess the most likely values of $X_1, X_2, \ldots$

Example: hidden Markov model



hidden variables $\qquad X_i \in \Sigma = \{\mathrm{in}, \mathrm{ex}\}$

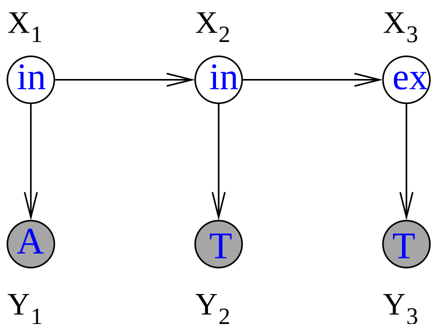
observed variables $\qquad Y_i \in \Sigma' = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$

Transition probabilities: for $i, j \in \Sigma$, $\ell \in \Sigma'$,

$$\left.\begin{aligned} \theta_{ij} &= \mathrm{Prob}(X_{k+1} = j \mid X_k = i) \\ \theta_{i\ell} &= \mathrm{Prob}(Y_k = \ell \mid X_k = i) \end{aligned}\right\} \text{12 parameters: } \theta_{\mathrm{in},\mathrm{in}},\ \theta_{\mathrm{in},\mathtt{A}}, \ldots$$

- The only information that we see are the values of $Y_1, Y_2, \ldots$. We want to guess the most likely values of $X_1, X_2, \ldots$
- This model can be used to determine what parts of the genome are introns and what parts are exons.

Example: hidden Markov model



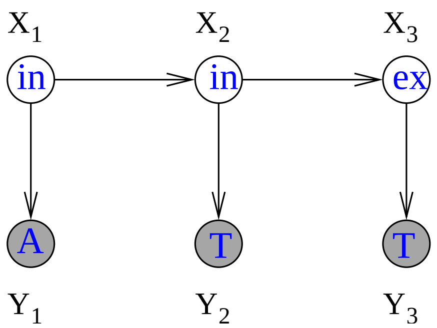
} hidden variables $\qquad X_i \in \Sigma = \{\mathrm{in}, \mathrm{ex}\}$

} observed variables $\qquad Y_i \in \Sigma' = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$

- The model is represented by a polynomial map

$$\begin{array}{rccc} \mathbf{f}: & \mathbb{R}^{12} & \longrightarrow & \mathbb{R}^{64} \\ & (\theta_{\mathrm{in},\mathrm{in}}, \theta_{\mathrm{in},\mathrm{ex}}, \theta_{\mathrm{in},\mathtt{A}}, \ldots) & \mapsto & (f_{\mathtt{AAA}}, f_{\mathtt{AAC}}, f_{\mathtt{AAG}}, f_{\mathtt{AAT}}, f_{\mathtt{ACA}}, \ldots) \end{array}$$

Example: hidden Markov model



} hidden variables $\qquad X_i \in \Sigma = \{\mathrm{in}, \mathrm{ex}\}$

} observed variables $\qquad Y_i \in \Sigma' = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$

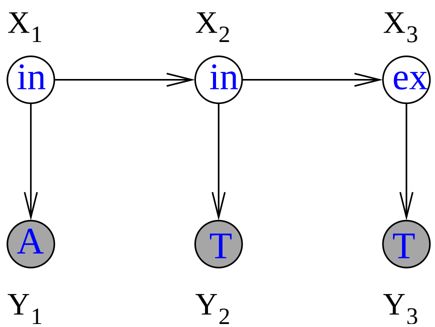
- The model is represented by a polynomial map

$$\begin{array}{rccc} \mathbf{f}: & \mathbb{R}^{12} & \longrightarrow & \mathbb{R}^{64} \\ & (\theta_{\mathrm{in},\mathrm{in}}, \theta_{\mathrm{in},\mathrm{ex}}, \theta_{\mathrm{in},\mathtt{A}}, \ldots) & \mapsto & (f_{\mathtt{AAA}}, f_{\mathtt{AAC}}, f_{\mathtt{AAG}}, f_{\mathtt{AAT}}, f_{\mathtt{ACA}}, \ldots) \end{array}$$

$$f_{\mathtt{ATT}} = \underbrace{\frac{1}{2}\theta_{\mathrm{in},\mathtt{A}}\theta_{\mathrm{in},\mathrm{in}}\theta_{\mathrm{in},\mathtt{T}}\theta_{\mathrm{in},\mathrm{in}}\theta_{\mathrm{in},\mathtt{T}}}_{X_1=X_2=X_3=\mathrm{in}} + \underbrace{\frac{1}{2}\theta_{\mathrm{in},\mathtt{A}}\theta_{\mathrm{in},\mathrm{in}}\theta_{\mathrm{in},\mathtt{T}}\theta_{\mathrm{in},\mathrm{ex}}\theta_{\mathrm{ex},\mathtt{T}}}_{X_1=X_2=\mathrm{in},\ X_3=\mathrm{ex}} + \cdots$$

(assuming $\mathrm{Prob}(X_1 = \mathrm{in}) = \mathrm{Prob}(X_1 = \mathrm{ex}) = \frac{1}{2}$)

Example: hidden Markov model



} hidden variables $\qquad X_i \in \Sigma = \{\mathrm{in}, \mathrm{ex}\}$

} observed variables $\qquad Y_i \in \Sigma' = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$

- The model is represented by a polynomial map

$$\begin{array}{lccc} \mathbf{f}: & \mathbb{R}^{12} & \longrightarrow & \mathbb{R}^{64} \\ & (\theta_{\mathrm{in,in}}, \theta_{\mathrm{in,ex}}, \theta_{\mathrm{in,\mathtt{A}}}, \ldots) & \mapsto & (f_{\mathtt{AAA}}, f_{\mathtt{AAC}}, f_{\mathtt{AAG}}, f_{\mathtt{AAT}}, f_{\mathtt{ACA}}, \ldots) \end{array}$$

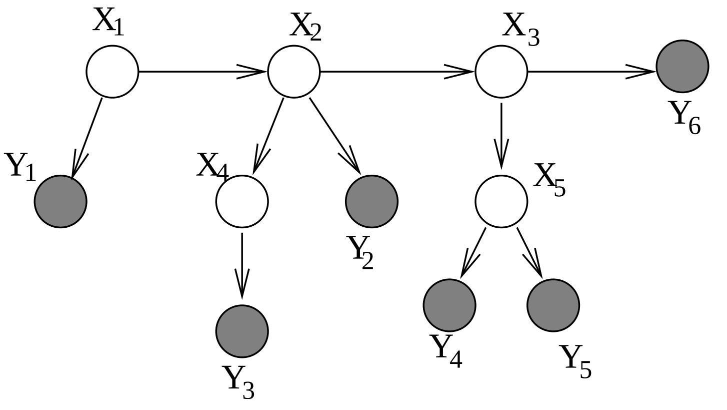
$$f_{\mathtt{ATT}} = \underbrace{\frac{1}{2}\theta_{\mathrm{in,\mathtt{A}}}\theta_{\mathrm{in,in}}\theta_{\mathrm{in,\mathtt{T}}}\theta_{\mathrm{in,in}}\theta_{\mathrm{in,\mathtt{T}}}}_{X_1=X_2=X_3=\mathrm{in}} + \underbrace{\frac{1}{2}\theta_{\mathrm{in,\mathtt{A}}}\theta_{\mathrm{in,in}}\theta_{\mathrm{in,\mathtt{T}}}\theta_{\mathrm{in,ex}}\theta_{\mathrm{ex,\mathtt{T}}}}_{X_1=X_2=\mathrm{in},\ X_3=\mathrm{ex}} + \cdots$$

(assuming $\mathrm{Prob}(X_1 = \mathrm{in}) = \mathrm{Prob}(X_1 = \mathrm{ex}) = \frac{1}{2}$)

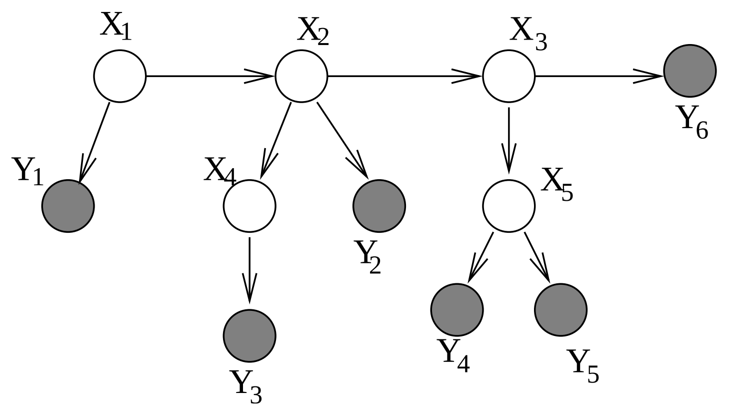
- Given an observation (e.g. $\mathbf{Y} = \mathtt{ATT}$), we want to find the most likely values of X_1, X_2, X_3.

Graphical models

More generally:

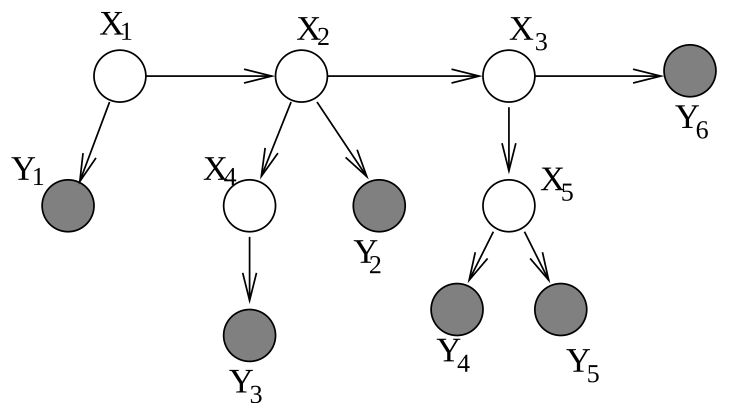


More generally:



- Hidden random variables: $X_1, X_2, \ldots, X_q \in \Sigma$
- Observed random variables: $Y_1, Y_2, \ldots, Y_n \in \Sigma'$

More generally:



- Hidden random variables: $X_1, X_2, \ldots, X_q \ \in \Sigma$
- Observed random variables: $Y_1, Y_2, \ldots, Y_n \ \in \Sigma'$

- Transition probabilities among states are given by parameters $\theta_1, \theta_2, \ldots, \theta_d$.

The coordinates of the map $\mathbf{f} : \mathbb{R}^d \longrightarrow \mathbb{R}^{(l')^n}$ are

$$f_{\tau} = \mathrm{Prob}(\mathbf{Y} = \tau) = \sum_{\mathbf{h} \in \Sigma^q} \underbrace{\mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau)}_{\text{monomial in } \theta_1, \theta_2, \ldots, \theta_d}$$

for each observation $\tau \in (\Sigma')^n$.

The coordinates of the map $\mathbf{f} : \mathbb{R}^d \longrightarrow \mathbb{R}^{(l')^n}$ are

$$f_\tau = \mathrm{Prob}(\mathbf{Y} = \tau) = \sum_{\mathbf{h} \in \Sigma^q} \underbrace{\mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau)}_{\text{monomial in } \theta_1, \theta_2, \ldots, \theta_d}$$

for each observation $\tau \in (\Sigma')^n$.

Basic inference problem: Given an observation τ, determine the value of the hidden data $\mathbf{h}$ that maximizes

$$\mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau).$$

The coordinates of the map $\mathbf{f} : \mathbb{R}^d \longrightarrow \mathbb{R}^{(l')^n}$ are

$$f_\tau = \mathrm{Prob}(\mathbf{Y} = \tau) = \sum_{\mathbf{h} \in \Sigma^q} \underbrace{\mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau)}_{\text{monomial in } \theta_1, \theta_2, \ldots, \theta_d}$$

for each observation $\tau \in (\Sigma')^n$.

Basic inference problem: Given an observation τ, determine the value of the hidden data $\mathbf{h}$ that maximizes

$$\mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau).$$

- The solution is denoted $\widehat{\mathbf{h}}$ and it is called an *explanation* of the observation τ.

The coordinates of the map $\mathbf{f} : \mathbb{R}^d \longrightarrow \mathbb{R}^{(l')^n}$ are

$$f_\tau = \mathrm{Prob}(\mathbf{Y} = \tau) = \sum_{\mathbf{h} \in \Sigma^q} \underbrace{\mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau)}_{\text{monomial in } \theta_1, \theta_2, \ldots, \theta_d}$$

for each observation $\tau \in (\Sigma')^n$.

Basic inference problem: Given an observation τ, determine the value of the hidden data $\mathbf{h}$ that maximizes

$$\mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau).$$

- The solution is denoted $\widehat{\mathbf{h}}$ and it is called an *explanation* of the observation τ.
- The explanation depends on the choice of the parameters $\theta_1, \theta_2, \ldots, \theta_d$.

Definition. An *inference function* is a map

$$\begin{array}{ccc} (\Sigma')^n & \longrightarrow & \Sigma^q \\ \tau & \mapsto & \widehat{\mathbf{h}} \end{array}$$

that sends each observation to its explanation.

Inference functions

Definition. An *inference function* is a map

$$\begin{array}{ccc} (\Sigma')^n & \longrightarrow & \Sigma^q \\ \tau & \mapsto & \widehat{\mathbf{h}} \end{array}$$

that sends each observation to its explanation.

- Each choice of parameters $(\theta_1, \theta_2, \ldots, \theta_d)$ defines an inference function.

Definition. An *inference function* is a map

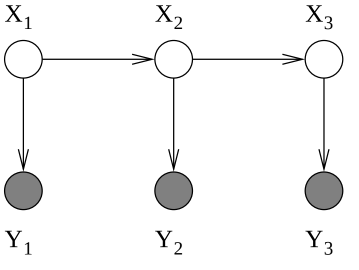
$$\begin{array}{ccc} (\Sigma')^n & \longrightarrow & \Sigma^q \\ \tau & \mapsto & \widehat{\mathbf{h}} \end{array}$$

that sends each observation to its explanation.

- Each choice of parameters $(\theta_1, \theta_2, \ldots, \theta_d)$ defines an inference function.

Example: Inference functions of the previous HMM are *gene finding functions*

$$\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^n \longrightarrow \{\mathrm{in}, \mathrm{ex}\}^n.$$



$$\Sigma = \{\mathrm{in}, \mathrm{ex}\}$$

$$\Sigma' = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$$

In general, if $l = |\Sigma|$, $l' = |\Sigma'|$, there are in total $l^{q(l')^n}$ functions

$$(\Sigma')^n \longrightarrow \Sigma^q.$$

The number of inference functions

In general, if $l = |\Sigma|$, $l' = |\Sigma'|$, there are in total $l^{q(l')^n}$ functions

$$(\Sigma')^n \longrightarrow \Sigma^q.$$

However, most of these maps are not inference functions for any value of the parameters.

The number of inference functions

In general, if $l = |\Sigma|$, $l' = |\Sigma'|$, there are in total $l^{q(l')^n}$ functions

$$(\Sigma')^n \longrightarrow \Sigma^q.$$

However, most of these maps are not inference functions for any value of the parameters.

Problem: How many of them are inference functions?

The number of inference functions

In general, if $l = |\Sigma|$, $l' = |\Sigma'|$, there are in total $l^{q(l')^n}$ functions

$$(\Sigma')^n \longrightarrow \Sigma^q.$$

However, most of these maps are not inference functions for any value of the parameters.

Problem: How many of them are inference functions?

Main Theorem. Fix $d > 0$. Consider a graphical model with d parameters, and let E be the number of edges of the underlying graph. Then, the number of inference functions of the model is at most

$$O(E^{d(d-1)}).$$

In general, if $l = |\Sigma|$, $l' = |\Sigma'|$, there are in total $l^{q(l')^n}$ functions

$$(\Sigma')^n \longrightarrow \Sigma^q.$$

However, most of these maps are not inference functions for any value of the parameters.

Problem: How many of them are inference functions?

Main Theorem. Fix $d > 0$. Consider a graphical model with d parameters, and let E be the number of edges of the underlying graph. Then, the number of inference functions of the model is at most

$$O(E^{d(d-1)}).$$

In most graphical models, E is a linear function of n, so the number of inference functions is $O(n^{d(d-1)})$.
For example, in the HMM, $E = 2n - 1$.

Geometric interpretation of the inference problem

For fixed parameters and a given observation τ, the inference problem is equivalent to find:

$$\boxed{\text{The explanation } \widehat{\mathbf{h}} \text{ that maximizes } \mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau).}$$

Geometric interpretation of the inference problem

For fixed parameters and a given observation τ, the inference problem is equivalent to find:

$$\boxed{\text{The explanation } \widehat{\mathbf{h}} \text{ that maximizes } \mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau).}$$

$$\Big\updownarrow \qquad f_\tau = \sum_{\mathbf{h} \in \Sigma^q} \mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau)$$

$$\boxed{\text{The monomial } \theta_1^{a_{1,i}} \theta_2^{a_{2,i}} \cdots \theta_d^{a_{d,i}} \text{ of } f_\tau \text{ with maximum value.}}$$

Geometric interpretation of the inference problem

For fixed parameters and a given observation τ, the inference problem is equivalent to find:

$\boxed{\text{The explanation } \widehat{\mathbf{h}} \text{ that maximizes } \mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau).}$

$$\Big\updownarrow \qquad f_\tau = \sum_{\mathbf{h} \in \Sigma^q} \mathrm{Prob}(\mathbf{X} = \mathbf{h}, \mathbf{Y} = \tau)$$

$\boxed{\text{The monomial } \theta_1^{a_{1,i}} \theta_2^{a_{2,i}} \cdots \theta_d^{a_{d,i}} \text{ of } f_\tau \text{ with maximum value.}}$

$$\updownarrow \qquad v_i := \log(\theta_i)$$

The monomial of f_τ whose exponent vector $(a_{1,i}, a_{2,i}, \ldots, a_{d,i})$ maximizes the linear expression $v_1 x_1 + \ldots + v_d x_d$.

Definition. The *Newton polytope* of

$$f_{\tau}(\theta_1, \theta_2, \ldots, \theta_d) = \sum_i \theta_1^{a_{1,i}} \theta_2^{a_{2,i}} \cdots \theta_d^{a_{d,i}},$$

$\mathrm{NP}(f_{\tau})$, is the convex hull of the vectors $(a_{1,i}, a_{2,i}, \ldots, a_{d,i}) \in \mathbb{R}^d$.

Definition. The *Newton polytope* of

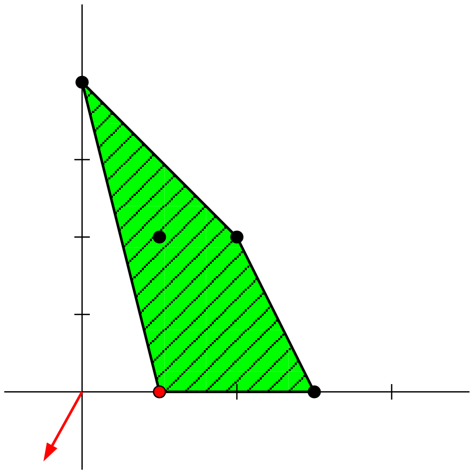
$$f_\tau(\theta_1, \theta_2, \ldots, \theta_d) = \sum_i \theta_1^{a_{1,i}} \theta_2^{a_{2,i}} \cdots \theta_d^{a_{d,i}},$$

$\mathrm{NP}(f_\tau)$, is the convex hull of the vectors $(a_{1,i}, a_{2,i}, \ldots, a_{d,i}) \in \mathbb{R}^d$.

Example.

$$f(\theta_1, \theta_2) = \theta_1^3 + \theta_1^2\theta_2^2 + \theta_1\theta_2^2 + \theta_1 + \theta_2^4$$

$$\theta_1 = 0.6, \quad \theta_2 = 0.4,$$
$$\mathrm{v} = (\log\theta_1, \log\theta_2) = (-0.51, -0.92)$$



Definition. The *Newton polytope* of

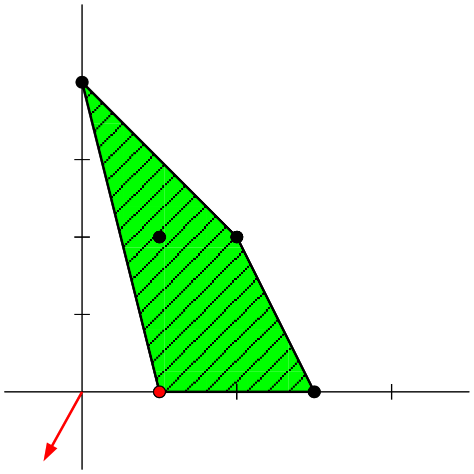
$$f_\tau(\theta_1, \theta_2, \ldots, \theta_d) = \sum_i \theta_1^{a_{1,i}} \theta_2^{a_{2,i}} \cdots \theta_d^{a_{d,i}},$$

$\mathrm{NP}(f_\tau)$, is the convex hull of the vectors $(a_{1,i}, a_{2,i}, \ldots, a_{d,i}) \in \mathbb{R}^d$.

Example.

$$f(\theta_1, \theta_2) = \theta_1^3 + \theta_1^2\theta_2^2 + \theta_1\theta_2^2 + \theta_1 + \theta_2^4$$

$$\theta_1 = 0.6, \quad \theta_2 = 0.4,$$
$$\mathrm{v} = (\log\theta_1, \log\theta_2) = (-0.51, -0.92)$$



- The explanation of τ is given by the vertex of $\mathrm{NP}(f_\tau)$ that maximizes $v_1 x_1 + \ldots + v_d x_d$.

Definition. The *Newton polytope* of

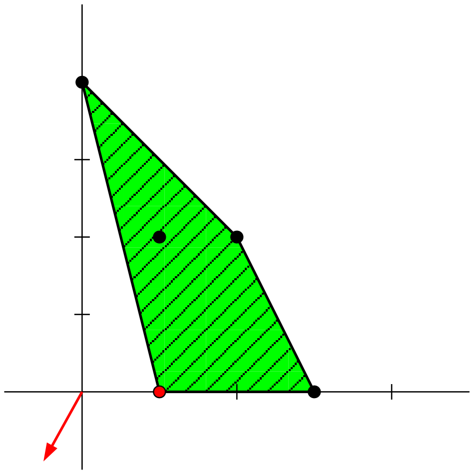
$$f_{\tau}(\theta_1, \theta_2, \ldots, \theta_d) = \sum_i \theta_1^{a_{1,i}} \theta_2^{a_{2,i}} \cdots \theta_d^{a_{d,i}},$$

$\mathrm{NP}(f_{\tau})$, is the convex hull of the vectors $(a_{1,i}, a_{2,i}, \ldots, a_{d,i}) \in \mathbb{R}^d$.

Example.

$$f(\theta_1, \theta_2) = \theta_1^3 + \theta_1^2\theta_2^2 + \theta_1\theta_2^2 + \theta_1 + \theta_2^4$$

$$\theta_1 = 0.6, \quad \theta_2 = 0.4,$$
$$\mathrm{v} = (\log\theta_1, \log\theta_2) = (-0.51, -0.92)$$



- The explanation of τ is given by the vertex of $\mathrm{NP}(f_{\tau})$ that maximizes $v_1x_1 + \ldots + v_dx_d$.

The inference problem for fixed parameters is a linear programming problem.

- Let
$$\Phi_{\mathbf{v}} : (\Sigma')^n \longrightarrow \Sigma^q$$
denote the inference function where $\mathbf{v} = (\log(\theta_1), \ldots, \log(\theta_d)) \in \mathbb{R}^d$.

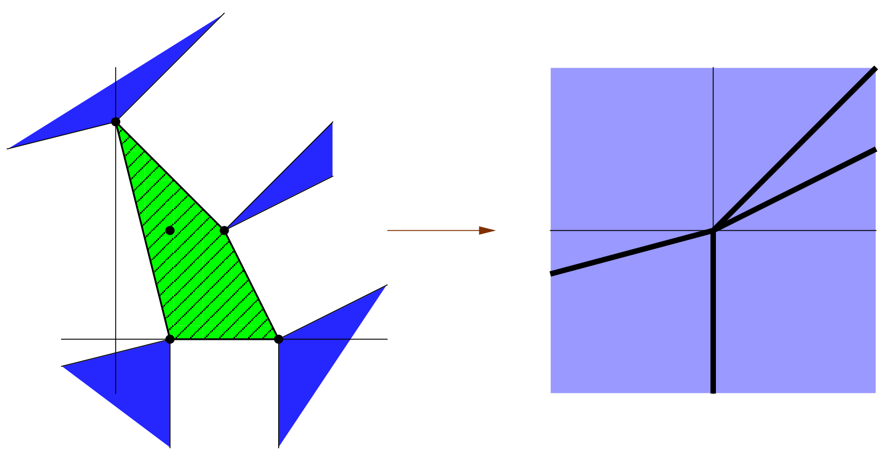
- Let
$$\Phi_{\mathbf{v}} : (\Sigma')^n \longrightarrow \Sigma^q$$
denote the inference function where $\mathbf{v} = (\log(\theta_1), \ldots, \log(\theta_d)) \in \mathbb{R}^d$.

- $\Phi_{\mathbf{v}}(\tau)$ is given by the vertex of $\mathrm{NP}(f_\tau)$ that is maximal in the direction of the vector $\mathbf{v}$.

The normal fan

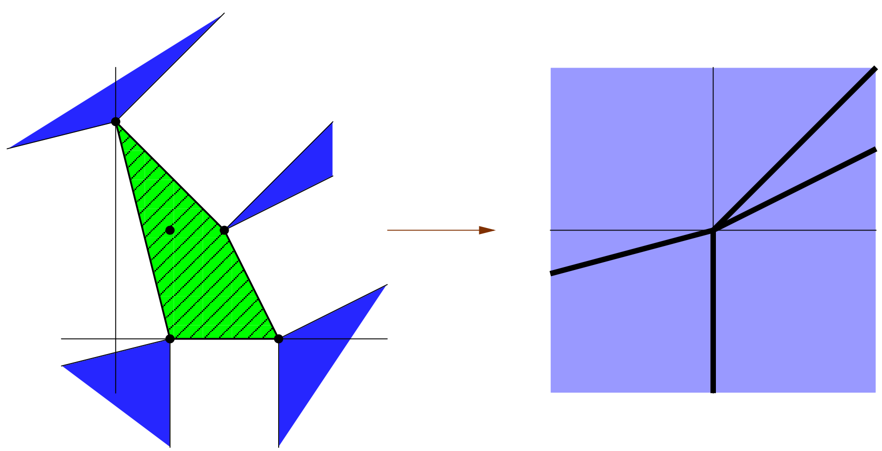
- Let $$\Phi_{\mathbf{v}} : (\Sigma')^n \longrightarrow \Sigma^q$$ denote the inference function where $\mathbf{v} = (\log(\theta_1), \ldots, \log(\theta_d)) \in \mathbb{R}^d$.
- $\Phi_{\mathbf{v}}(\tau)$ is given by the vertex of $\mathrm{NP}(f_\tau)$ that is maximal in the direction of the vector $\mathbf{v}$.

Definition. The *normal fan* $\mathcal{F}(P)$ of a polytope P is the decomposition of the space according to the directions in which each face is maximal.



- Let $$\Phi_{\mathbf{v}} : (\Sigma')^n \longrightarrow \Sigma^q$$
  denote the inference function where $\mathbf{v} = (\log(\theta_1), \ldots, \log(\theta_d)) \in \mathbb{R}^d$.
- $\Phi_{\mathbf{v}}(\tau)$ is given by the vertex of $\mathrm{NP}(f_\tau)$ that is maximal in the direction of the vector $\mathbf{v}$.

Definition. The *normal fan* $\mathcal{F}(P)$ of a polytope P is the decomposition of the space according to the directions in which each face is maximal.



$\Phi_{\mathbf{v}}(\tau) = \Phi_{\mathbf{v}'}(\tau)$ iff $\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\mathcal{F}(\mathrm{NP}(f_\tau))$.

Vectors and inference functions

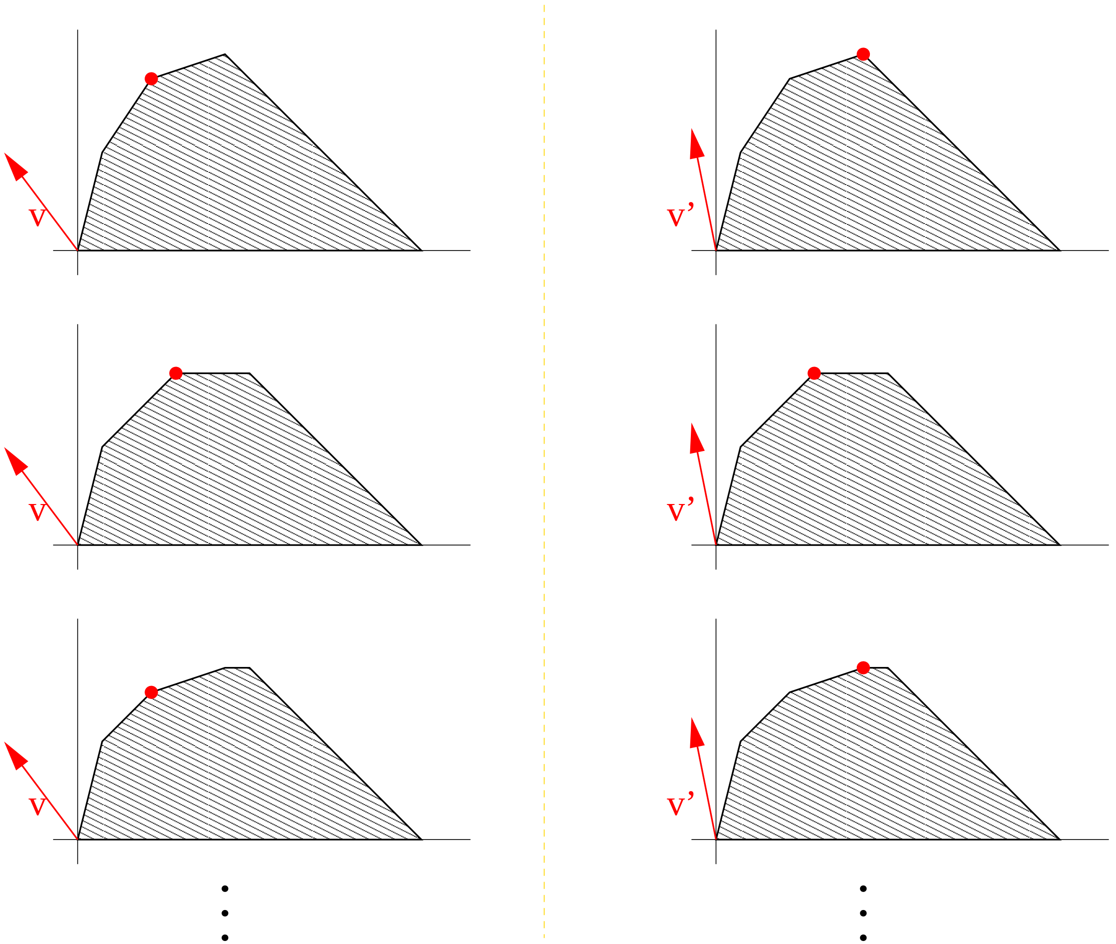
Different choices of parameters may determine different inference functions.

How many different inference functions can a graphical model have?

Vectors and inference functions

Different choices of parameters may determine different inference functions.

How many different inference functions can a graphical model have?



Common refinement of fans

$$\Phi_{\mathbf{v}}(\tau) = \Phi_{\mathbf{v}'}(\tau) \quad \Longleftrightarrow \quad \begin{array}{c}\mathbf{v} \text{ and } \mathbf{v}' \text{ belong to the} \\ \text{same cone of } \mathcal{F}(\mathrm{NP}(f_\tau))\end{array}$$

Common refinement of fans

$\Phi_{\mathbf{v}}(\tau) = \Phi_{\mathbf{v}'}(\tau) \quad \Longleftrightarrow$ $\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\mathcal{F}(\mathrm{NP}(f_\tau))$

$\Phi_{\mathbf{v}} \equiv \Phi_{\mathbf{v}'}$ as inference functions $\Longleftrightarrow$ $\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\mathcal{F}(\mathrm{NP}(f_\tau))$ for all observations $\tau \in (\Sigma')^n$

Common refinement of fans

$\Phi_{\mathbf{v}}(\tau) = \Phi_{\mathbf{v}'}(\tau) \iff$ $\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\mathcal{F}(\mathrm{NP}(f_\tau))$

$\Phi_{\mathbf{v}} \equiv \Phi_{\mathbf{v}'}$ as inference functions $\iff$ $\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\mathcal{F}(\mathrm{NP}(f_\tau))$ for all observations $\tau \in (\Sigma')^n$

$\Updownarrow$

$\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\underbrace{\bigwedge_{\tau \in (\Sigma')^n} \mathcal{F}(\mathrm{NP}(f_\tau))}_{\text{common refinement of fans}}$

Common refinement of fans

$\Phi_{\mathbf{v}}(\tau) = \Phi_{\mathbf{v}'}(\tau) \quad \Longleftrightarrow$ $\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\mathcal{F}(\mathrm{NP}(f_\tau))$

$\Phi_{\mathbf{v}} \equiv \Phi_{\mathbf{v}'}$ as inference functions $\Longleftrightarrow$ $\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\mathcal{F}(\mathrm{NP}(f_\tau))$ for all observations $\tau \in (\Sigma')^n$

$\Updownarrow$

$\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\underbrace{\bigwedge_{\tau \in (\Sigma')^n} \mathcal{F}(\mathrm{NP}(f_\tau))}_{\text{common refinement of fans}}$

The number of inference functions equals the **number of cones in the common refinement of fans.**

- The common refinement of the normal fans is the normal fan of the Minkowski sum of polytopes:

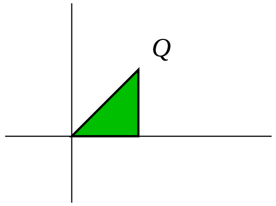
$$\bigwedge_{\tau \in (\Sigma')^n} \mathcal{F}(\mathrm{NP}(f_\tau)) = \mathcal{F}\Big(\bigodot_{\tau \in (\Sigma')^n} \mathrm{NP}(f_\tau)\Big)$$

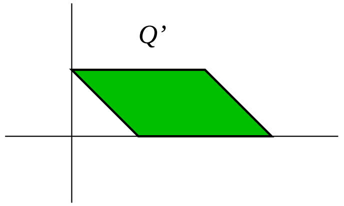
Minkowski sum of polytopes

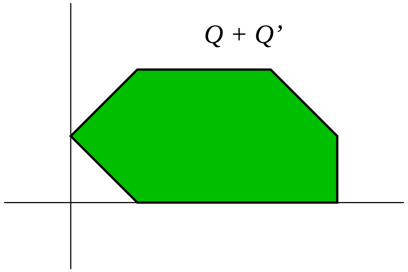
- The common refinement of the normal fans is the normal fan of the Minkowski sum of polytopes:

$$\bigwedge_{\tau \in (\Sigma')^n} \mathcal{F}(\mathrm{NP}(f_\tau)) = \mathcal{F}\Big(\bigodot_{\tau \in (\Sigma')^n} \mathrm{NP}(f_\tau)\Big)$$

Definition. *Minkowski sum*: $Q \odot Q' := \{\mathbf{x} + \mathbf{x}' : \mathbf{x} \in Q,\, \mathbf{x}' \in Q'\}$.



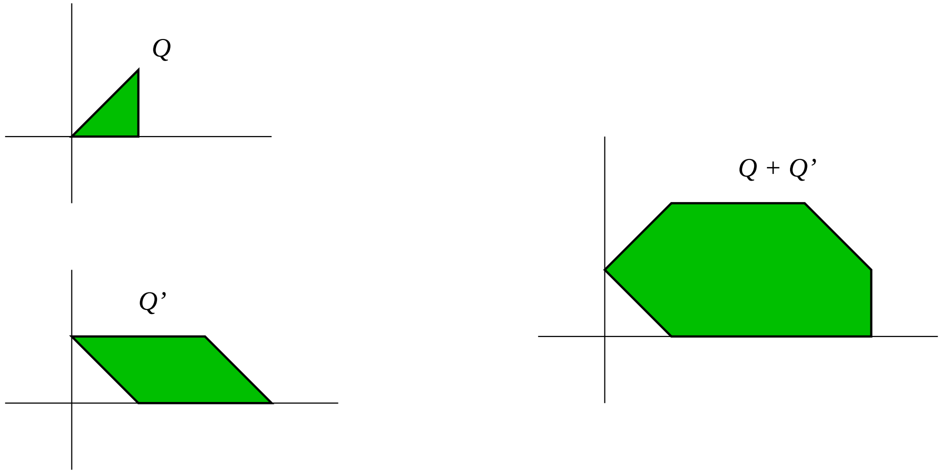




- The common refinement of the normal fans is the normal fan of the Minkowski sum of polytopes:

$$\bigwedge_{\tau \in (\Sigma')^n} \mathcal{F}(\mathrm{NP}(f_\tau)) = \mathcal{F}\Big(\bigodot_{\tau \in (\Sigma')^n} \mathrm{NP}(f_\tau) \Big)$$

Definition. *Minkowski sum*: $Q \odot Q' := \{\mathbf{x} + \mathbf{x}' : \mathbf{x} \in Q,\ \mathbf{x}' \in Q'\}$.

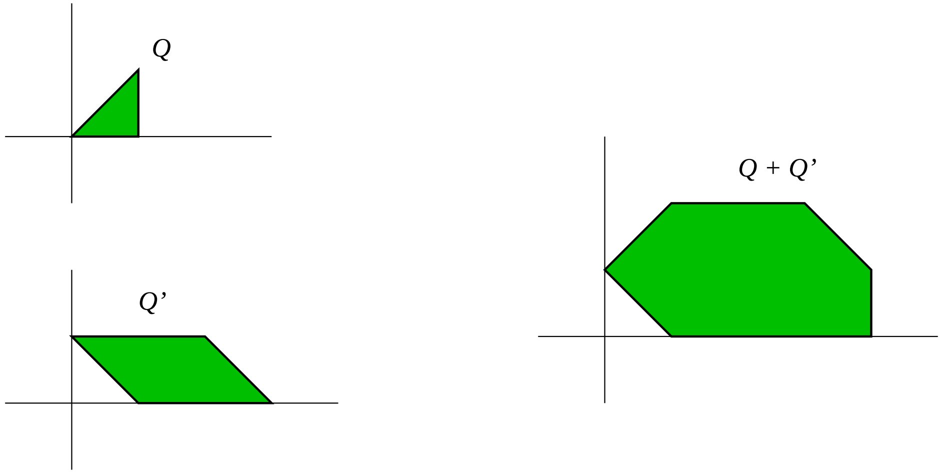


Let $P := \bigodot_{\tau \in (\Sigma')^n} \mathrm{NP}(f_\tau)$.

- The common refinement of the normal fans is the normal fan of the Minkowski sum of polytopes:

$$\bigwedge_{\tau \in (\Sigma')^n} \mathcal{F}(\mathrm{NP}(f_\tau)) = \mathcal{F}\Big(\bigodot_{\tau \in (\Sigma')^n} \mathrm{NP}(f_\tau)\Big)$$

Definition. *Minkowski sum*: $Q \odot Q' := \{\mathbf{x} + \mathbf{x}' : \mathbf{x} \in Q,\ \mathbf{x}' \in Q'\}$.



Let $P := \bigodot_{\tau \in (\Sigma')^n} \mathrm{NP}(f_\tau)$.

The number of inference functions equals the **number of vertices of** P.

- For each τ,

$$\mathrm{NP}(f_\tau) \subseteq [0, E]^d,$$

since each θ_i can appear as a factor of a monomial of f_τ at most E times.

- For each τ,

$$\mathrm{NP}(f_\tau) \subseteq [0, E]^d,$$

since each θ_i can appear as a factor of a monomial of f_τ at most E times.

- Besides, $\mathrm{NP}(f_\tau)$ is a lattice polytope (i.e., its vertices have integral coordinates).

- For each τ,

$$\mathrm{NP}(f_\tau) \subseteq [0, E]^d,$$

  since each θ_i can appear as a factor of a monomial of f_τ at most E times.

- Besides, $\mathrm{NP}(f_\tau)$ is a lattice polytope (i.e., its vertices have integral coordinates).

$$\Downarrow$$

Edges of $\mathrm{NP}(f_\tau)$ are vectors where each coordinate is an integer between $-E$ and E.

Edges of $\mathrm{NP}(f_\tau)$

- For each τ,

  $$\mathrm{NP}(f_\tau) \subseteq [0, E]^d,$$

  since each θ_i can appear as a factor of a monomial of f_τ at most E times.

- Besides, $\mathrm{NP}(f_\tau)$ is a lattice polytope (i.e., its vertices have integral coordinates).

$$\Downarrow$$

Edges of $\mathrm{NP}(f_\tau)$ are vectors where each coordinate is an integer between $-E$ and E.

$$\Downarrow$$

The polytopes $\mathrm{NP}(f_\tau)$ have at most $(2E+1)^d$ nonparallel edges in total.

The polytopes $\mathrm{NP}(f_\tau)$ have at most $(2E+1)^d$ nonparallel edges in total.

$$\Downarrow \text{ [Gritzmann, Sturmfels]}$$

The number of vertices of $P = \bigodot_{\tau \in (\Sigma')^n} \mathrm{NP}(f_\tau)$ is at most

$$2 \sum_{j=0}^{d-1} \binom{(2E+1)^d - 1}{j}.$$

The polytopes $\mathrm{NP}(f_\tau)$ have at most $(2E+1)^d$ nonparallel edges in total.

$$\Downarrow \text{[Gritzmann, Sturmfels]}$$

The number of vertices of $P = \bigodot_{\tau \in (\Sigma')^n} \mathrm{NP}(f_\tau)$ is at most

$$2 \sum_{j=0}^{d-1} \binom{(2E+1)^d - 1}{j}.$$

As $E \to \infty$, the dominant term is $\frac{2^{d^2-d+1}}{(d-1)!} E^{d(d-1)}$.

The polytopes $\mathrm{NP}(f_\tau)$ have at most $(2E+1)^d$ nonparallel edges in total.

$$\Downarrow \text{[Gritzmann, Sturmfels]}$$

The number of vertices of $P = \bigodot_{\tau \in (\Sigma')^n} \mathrm{NP}(f_\tau)$ is at most

$$2 \sum_{j=0}^{d-1} \binom{(2E+1)^d - 1}{j}.$$

As $E \to \infty$, the dominant term is $\frac{2^{d^2-d+1}}{(d-1)!} E^{d(d-1)}$.

$$\Downarrow$$

The number of inference functions is $O(E^{d(d-1)})$. $\square$

Application to sequence alignment

- The 2-parameter model for sequence alignment is a particular case of a *pair hidden Markov model*.

Application to sequence alignment

- The 2-parameter model for sequence alignment is a particular case of a *pair hidden Markov model*.

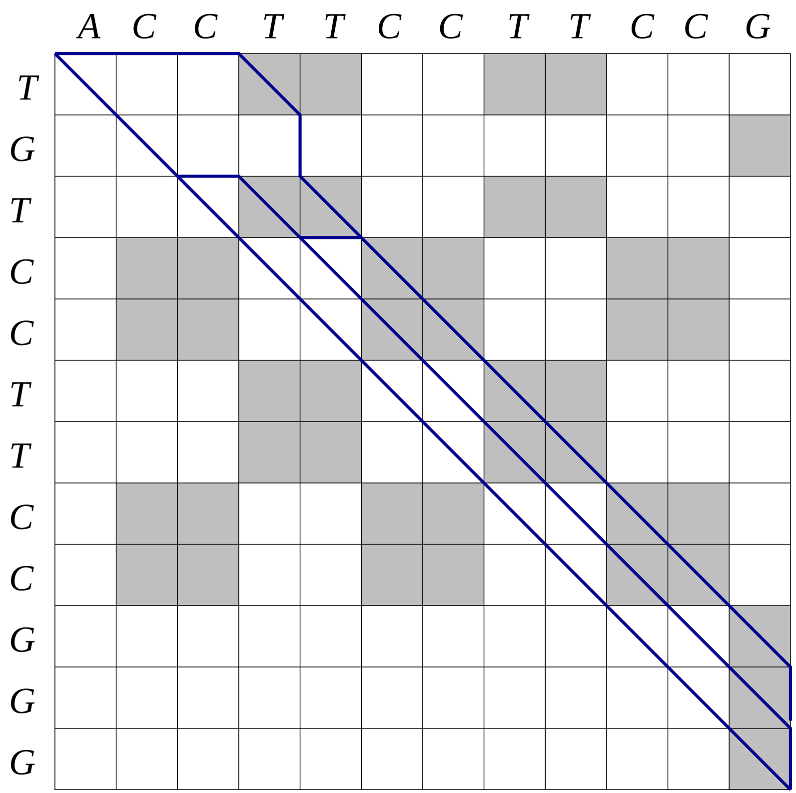
General graphical model		**Sequence alignment**
observation τ	$\longleftrightarrow$	pair of sequences of length n
explanation $\widehat{\mathbf{h}}$	$\longleftrightarrow$	optimal alignment

Application to sequence alignment

- The 2-parameter model for sequence alignment is a particular case of a *pair hidden Markov model*.

General graphical model		**Sequence alignment**
observation τ	$\longleftrightarrow$	pair of sequences of length n
explanation $\widehat{\mathbf{h}}$	$\longleftrightarrow$	optimal alignment
Newton polytope $\mathrm{NP}(f_\tau)$	$\longleftrightarrow$	convex hull of points (x, y, z) where $x = \#$mismatches, $y = \#$insertions, $z = \#$matches for each alignment (this polytope lies on the plane $x + y + z = n$)

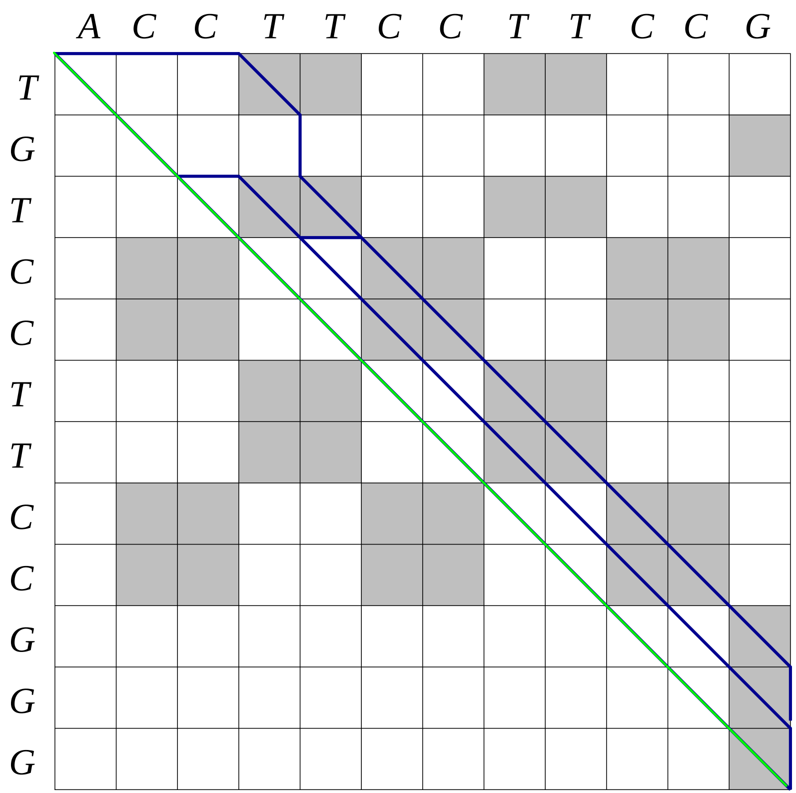
- Alignments can be represented as paths from the upper-left to the lower-right corner in the alignment graph:



x	y	z

Alignment graphs

- Alignments can be represented as paths from the upper-left to the lower-right corner in the alignment graph:



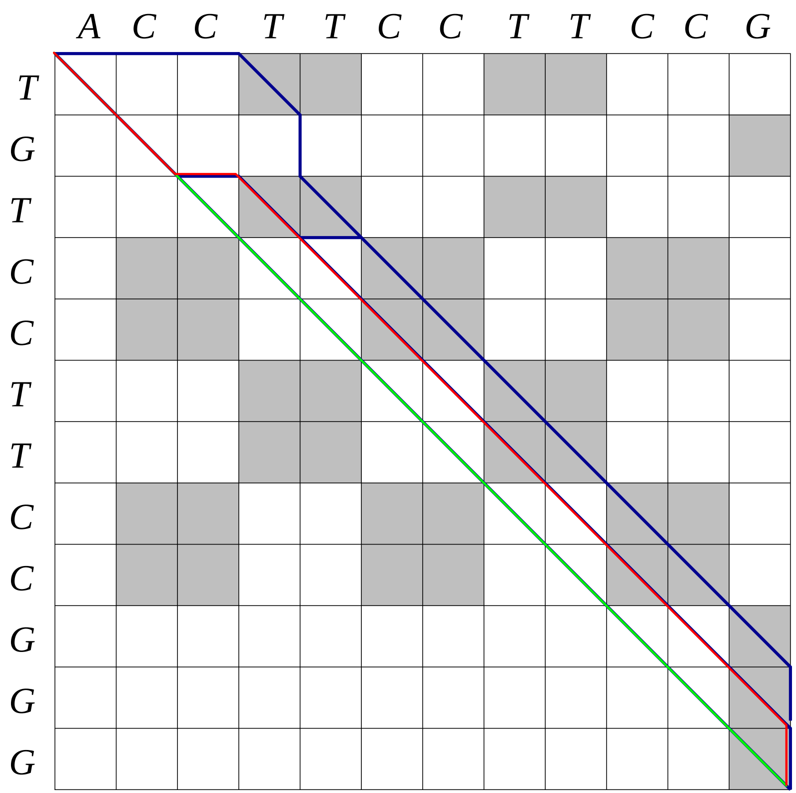
x	y	z
11	0	1

TGTCCTTCCGGG

ACCTTCCTTCCG

Alignment graphs

- Alignments can be represented as paths from the upper-left to the lower-right corner in the alignment graph:

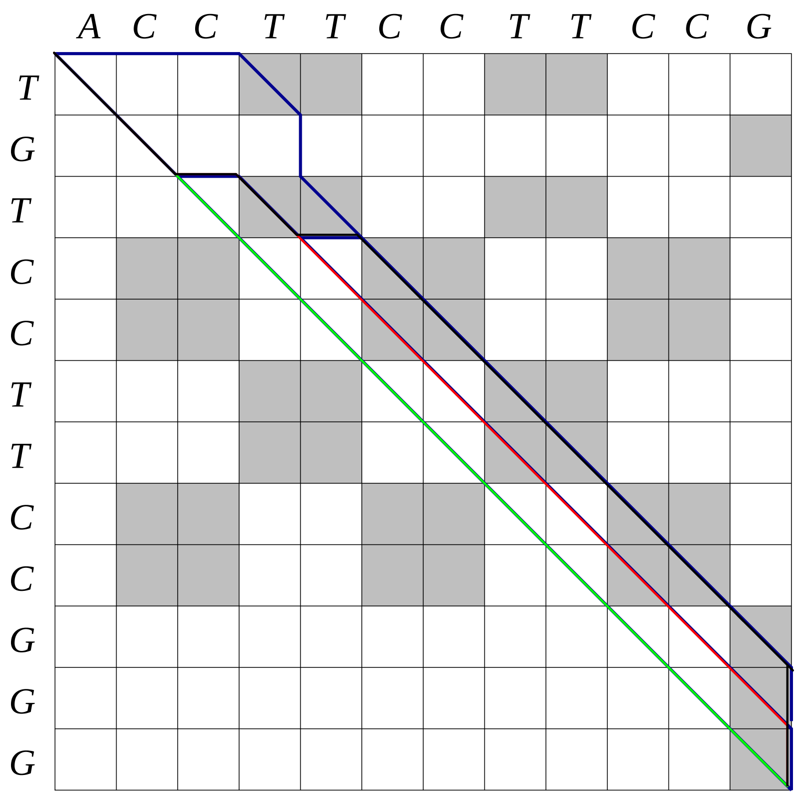


x	y	z
11	0	1
6	1	5

```
TGTCCTTCCGGG    TG-TCCTTCCGGG
ACCTTCCTTCCG    ACCTTCCTTCCG-
```

Alignment graphs

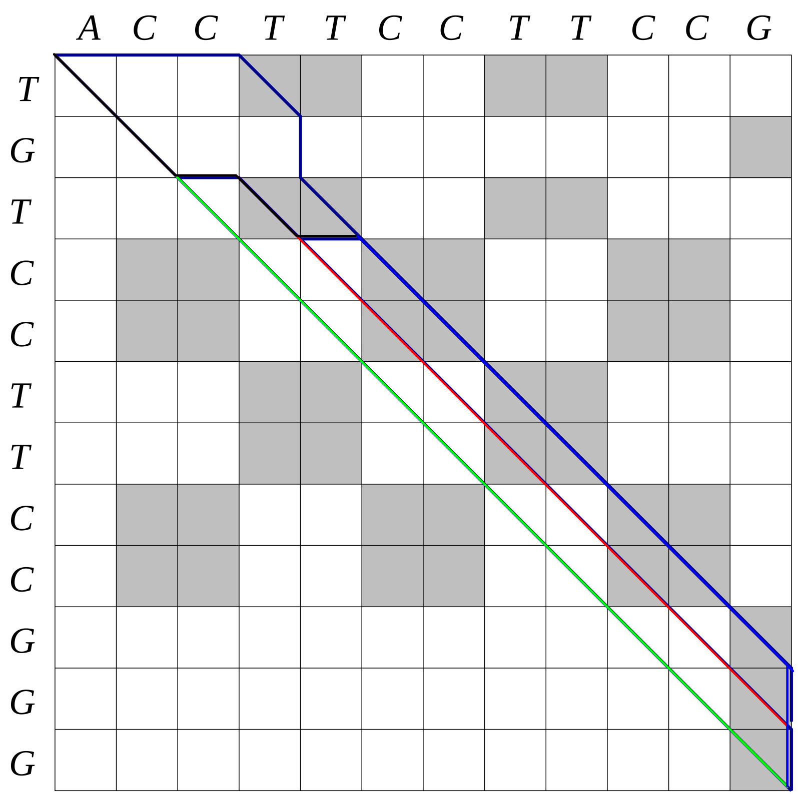
- Alignments can be represented as paths from the upper-left to the lower-right corner in the alignment graph:



x	y	z
11	0	1
6	1	5
2	2	8

```
TGTCCTTCCGGG    TG-TCCTTCCGGG    TG-T-CCTTCCGGG
ACCTTCCTTCCG    ACCTTCCTTCCG-    ACCTTCCTTCCG--
```

- Alignments can be represented as paths from the upper-left to the lower-right corner in the alignment graph:



x	y	z
11	0	1
6	1	5
2	2	8
0	3	9
⋮	⋮	⋮

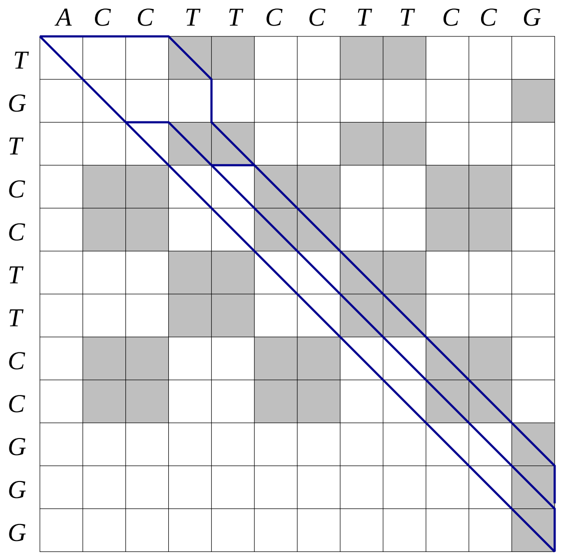
```
TGTCCTTCCGGG
ACCTTCCTTCCG
```

```
TG-TCCTTCCGGG
ACCTTCCTTCCG-
```

```
TG-T-CCTTCCGGG
ACCTTCCTTCCG--
```

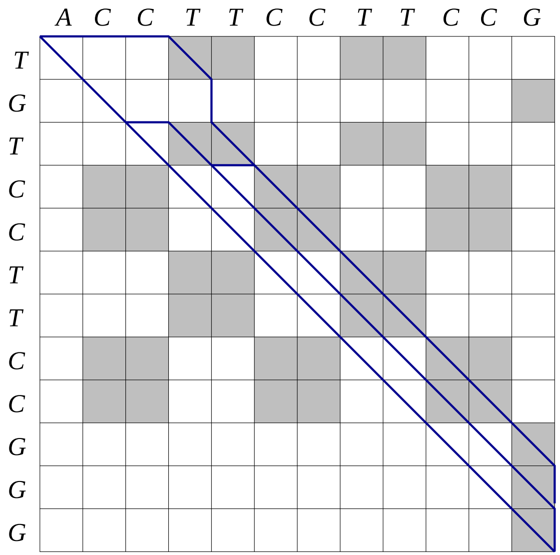
```
---TGTCCTTCCGGG
ACCT-TCCTTCCG--
```

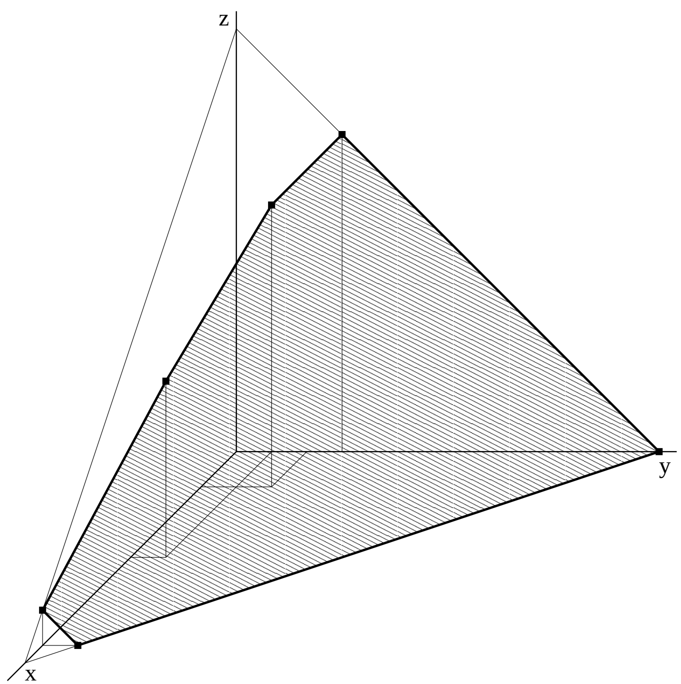
The Newton polytope of a pair of sequences



x	y	z
11	0	1
6	1	5
2	2	8
0	3	9
11	1	0
0	12	0

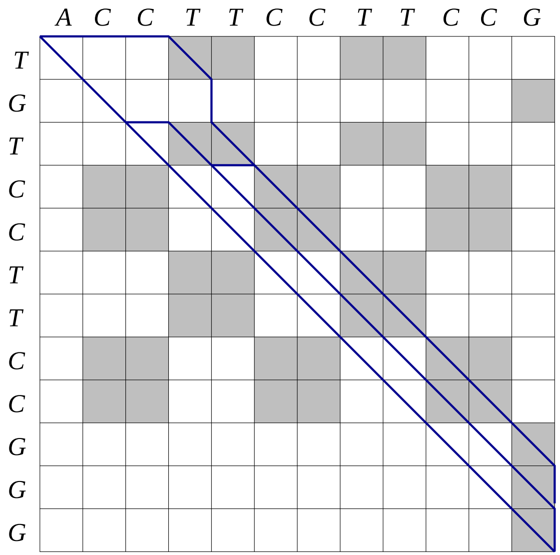
The Newton polytope of a pair of sequences



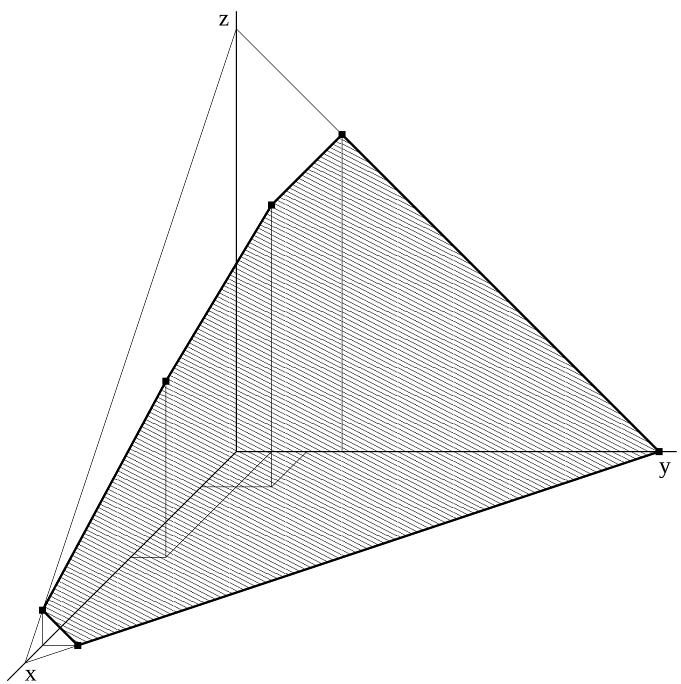


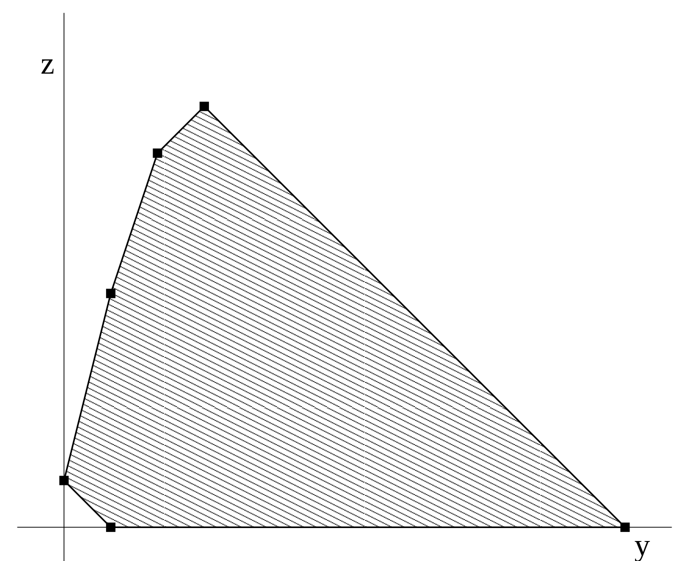
x	y	z
11	0	1
6	1	5
2	2	8
0	3	9
11	1	0
0	12	0

The Newton polytope of a pair of sequences



x	y	z
11	0	1
6	1	5
2	2	8
0	3	9
11	1	0
0	12	0





Bound on the number of inference functions

Let $\tau = (S, T)$ (sequences of length n).

In any alignment with $y = \#$insertions, $z = \#$matches, $(y, z) \in [0, n]^2$.

Bound on the number of inference functions

Let $\tau = (S, T)$ (sequences of length n).

In any alignment with $y = \#$insertions, $z = \#$matches, $(y, z) \in [0, n]^2$.

$$\Downarrow$$

(The projection of) $\mathrm{NP}(f_\tau)$ is a lattice polygon inside $[0, n]^2$.

Bound on the number of inference functions

Let $\tau = (S, T)$ (sequences of length n).

In any alignment with $y = \#$insertions, $z = \#$matches, $(y, z) \in [0, n]^2$.

$$\Downarrow$$

(The projection of) $\mathrm{NP}(f_\tau)$ is a lattice polygon inside $[0, n]^2$.

$\Downarrow$ [Main theorem]

The number of inference functions of this 2-parameter model is $O(n^{d(d-1)}) = O(n^2)$.

Bound on the number of inference functions

Let $\tau = (S, T)$ (sequences of length n).

In any alignment with $y = \#$insertions, $z = \#$matches, $(y, z) \in [0, n]^2$.

$$\Downarrow$$

(The projection of) $\mathrm{NP}(f_\tau)$ is a lattice polygon inside $[0, n]^2$.

$$\Downarrow \quad \text{[Main theorem]}$$

The number of inference functions of this 2-parameter model is $O(n^{d(d-1)}) = O(n^2)$.

It can be shown that this bound is tight, i.e.,

$$P = \bigodot_{\tau=(S,T)\in\{0,1\}^{2n}} \mathrm{NP}(f_\tau) \quad \text{has } \Theta(n^2) \text{ vertices.}$$

Open problems

- The bound of $O(E^{d(d-1)})$ on the number of inference functions is tight for $d = 2$.

  Open: Is this tight in higher dimensions? Can we get better bounds?

Open problems

- The bound of $O(E^{d(d-1)})$ on the number of inference functions is tight for $d = 2$.

  Open: Is this tight in higher dimensions? Can we get better bounds?

- In the 2-parameter model for sequence alignment,

  $$P = \bigodot_{\tau \in \{0,1\}^{2n}} \mathrm{NP}(f_\tau) \text{ has } \Theta(n^2) \text{ vertices } (= \# \text{ inference functions}).$$

  Open: For fixed $\tau = (S, T)$, how many vertices can $\mathrm{NP}(f_\tau)$ have?

Open problems

- The bound of $O(E^{d(d-1)})$ on the number of inference functions is tight for $d = 2$.

  Open: Is this tight in higher dimensions? Can we get better bounds?

- In the 2-parameter model for sequence alignment,

  $$P = \bigodot_{\tau \in \{0,1\}^{2n}} \mathrm{NP}(f_\tau) \text{ has } \Theta(n^2) \text{ vertices } (= \# \text{ inference functions}).$$

  Open: For fixed $\tau = (S, T)$, how many vertices can $\mathrm{NP}(f_\tau)$ have?

  This is the number of different alignments of S and T that can be optimal for some choice of parameters.

- The bound of $O(E^{d(d-1)})$ on the number of inference functions is tight for $d = 2$.

  Open: Is this tight in higher dimensions? Can we get better bounds?

- In the 2-parameter model for sequence alignment,

  $$P = \bigodot_{\tau \in \{0,1\}^{2n}} \mathrm{NP}(f_\tau) \text{ has } \Theta(n^2) \text{ vertices } (= \# \text{ inference functions}).$$

  Open: For fixed $\tau = (S, T)$, how many vertices can $\mathrm{NP}(f_\tau)$ have?

  This is the number of different alignments of S and T that can be optimal for some choice of parameters.

  Known to be $O(n^{2/3})$. [Gusfield et al '94, Fernández-Baca et al '02]

Open problems

- The bound of $O(E^{d(d-1)})$ on the number of inference functions is tight for $d = 2$.

  Open: Is this tight in higher dimensions? Can we get better bounds?

- In the 2-parameter model for sequence alignment,

  $$P = \bigodot_{\tau \in \{0,1\}^{2n}} \mathrm{NP}(f_\tau) \text{ has } \Theta(n^2) \text{ vertices } (= \# \text{ inference functions}).$$

  Open: For fixed $\tau = (S, T)$, how many vertices can $\mathrm{NP}(f_\tau)$ have?

  This is the number of different alignments of S and T that can be optimal for some choice of parameters.

  Known to be $O(n^{2/3})$. [Gusfield et al '94, Fernández-Baca et al '02]

  Conjecture: It is $O(\sqrt{n})$ for binary sequences.