

Multiple sequence alignment accuracy estimation and its role in creating an automated bioinformatician

Dan DeBlasio

dandeblasio.com



danfdeblasio


StringBio 2018



Computational
Biology
Department

Carnegie Mellon University
School of Computer Science

Tunable parameters



BioStars

BIOINFORMATICS EXPLAINED

[Community](#)[Log In](#)

Question: Salmon libtype - Does it matter?

Hi,

I was wondering what happens if you have the libtype wrong for Salmon? Or why it requires that I wasn't sure if the first read of my paired-end reads was forward (ISF) or reverse (ISR), so I ran it the mapping efficiencies were identical between the runs.

Thanks, Matt

salmon

alignme

ADD COMMENT • lib

Welcome to Biostar!

Community
Log In
Sign Up


Question: (Closed) Questions about Salmon

Hello All! I had some questions involving alignment based v alignment free mapping done by Salmon as well as some other general questions pertaining specifically to our experiments. Please excuse any perceived ignorance as I'm more of a molecular than computation biologist and statistics is not my strong suit!

0

- If one will probably map the reads in order to visualize downstream via tools like IGV, is it better to use the alignment based mapping so that Salmon agrees with the alignment output?
- Our experiments involve looking at ribosome profiling data when translation is disrupted - resulting in low read counts. We also have the corresponding RNA-Seq data. Should one use gcbias and seqbias for one/both data sets? My idea was to not use either option to avoid introducing bias into our already limited ribosome profiling data set.
- Could someone provide a better explanation for `fldMean` and `fldSD2`? We are using Illumina HiSeq 50 bp single end reads. Would


4 months ago by
[kyusikkim](#) • 10





BioStars

BIOINFORMATICS EXPLAINED


Welcome to BioStar!

 Community


 Log In

 Sign Up

Question: RNA-Seq analysis using STAR and Salmon



0



Hello! I am having some trouble figuring out how to use Salmon. I have around 30 different samples which I trimmed using bmap then aligned them using STAR. I have all of the BAM files from this alignment. Should I merge them all before running Salmon or because they are all unique samples, do I run them separately? I also aligned them to the UMD3.1 cattle genome. Is this ok for Salmon or should I align it to a different reference? I am very new to all this and trying to teach myself as I go. So if anyone has any other sites that could help me out, that would be great!

Thanks!

rna-seq


• 505 views

ADD COMMENT

• [link](#) • Not following


modified 4 months ago by [aka001](#) • 180

• written 4 months ago by [low m](#)



4 months ago by

[rachel.kubik12](#) • 0



Welcome to Biostar!

Community Log In Sign Up

Question: Salmon unmapped reads


Hi All

I have some samples with low mapping in Salmon (40% and less) that have higher alignments in Tophat, and trying to troubleshoot.

I picked some of the unmapped reads (from writeunmapped salmon parameter) and Blat them to human.

Some have 2 or more matches with identity 99% to 100% And some have many many matches, I need to scroll the page down too much. Many of these matches are 100% and some range between 85% to 100% identity.


I looked also into the "ambiguous" tag, found some records with 0 unique mapping and more than 100 ambiguous mapping, but couldn't



12 weeks ago by Sharon • 150

Mapping rate ? #160

biostars opened this issue on Oct 6, 2017. 7 comments



Biostars

BIOINFORMATICS EXPLAINED

Welcome to Biostar!

Community

Log In

Sign Up

Question: Salmon: Optimal k-mer size (for indexing) for RNA-seq data alignment using reference genome

1

Dear all,

I am using [salmon](#) to evaluate how the *L.donovani* gene expression varies in the two different (promastigotes and amastigotes). For that I downloaded two RNA-seq data from [SRA](#) and used reference *L.donovani* [coding sequence](#) coding sequences (CDS)

In order to quantify gene expression with salmon I have to index the CDS using a specific [salmon manual](#) they use a 31-mer for 75bp reads. My question is whether is reasonable to use this value i.e. 0,413reads length or if there's any other advisable value. I heard that some people use 31-mer length whereas I also found [this post](#) where is mentioned between 0,5 and 0,66*reads length rather for de novo genome assembly, which is not my case as I am mapping to a reference genome

Thanks in advance.

rna-seq

salmon

alignment

indexing

k-mer

• 1.2k views

ADD COMMENT

• [link](#) •

Not following ▼

modified 19 months ago


Question: Salmon very low mapping rate

0

Hi All

I am new to salmon. Got 26% mapping rate for my data. I have 100M reads, 75%. Although all the data have similar length, I am not sure if it is reasonable to use almost above 30 but with few bases at the end. I am not sure about how aggressive trimming is bad or good.

I also changed the k index during building the index, but I am not sure if it is over representative sequences and I am not sure if it is bad or good.



BioStars

BIOINFORMATICS EXPLAINED

[Community](#)[Log In](#)[Sign Up](#)

"Greater than 5% of the fragments disagreed with the provided issue. This is an example for one of the "lib_format_counts.json"


Question: Salmon very low mapping

Hi All

0

I am new to salmon. Got 26% mapping rate with one set and 43% with another. Rest of samples have more than 75%. Although all the data have similar quality and all have adapter content empty in FASTQC. The quality is almost above 30 but with few bases at the beginning less than 20 but greater than 10. I did not trim them, read about how aggressive trimming is bad.

I also changed the k index during building the index. Changed the reference transcript. Same results. I blat the over representative sequences and found no contaminants. I mapped the sample that has salmon 26% with Tophat default parameters. got 73% alignment. Not sure why?



12 weeks ago by
[Tania • 50](#)

Tunable parameters

Quant

=====

Perform dual-phase, mapping-based estimation of transcript abundance from RNA-seq reads

salmon quant options:

basic options:

-v [--version]	print version string
-h [--help]	produce help message
-i [--index] arg	Salmon index
-l [--libType] arg	Format string describing the library type
-r [--unmatedReads] arg	List of files containing unmated reads of (e.g. single-end reads)
-1 [--mates1] arg	File containing the #1 mates
-2 [--mates2] arg	File containing the #2 mates
-o [--output] arg	Output quantification file.
--discardOrphansQuasi	[Quasi-mapping mode only] : Discard orphan mappings in quasi-mapping mode. If this flag is passed then only paired mappings will be considered toward quantification estimates. The default behavior is to consider orphan mappings if no valid paired mappings exist. This flag is independent of the option to write the orphaned mappings to file (--writeOrphanLinks).
--allowOrphansFMD	[FMD-mapping mode only] : Consider orphaned reads as valid hits when performing lightweight-alignment. This option will increase sensitivity (allow more reads to map and more transcripts to be detected), but may decrease specificity as orphaned alignments are more likely to be spurious.
--seqBias	Perform sequence-specific bias correction.
--gcBias	[beta for single-end reads] Perform fragment GC bias correction
-p [--threads] arg	The number of threads to use concurrently.
--incompatPrior arg	This option sets the prior probability that an alignment that disagrees with the specified library type (--libType) results from the true fragment origin. Setting this to 0 specifies that alignments that disagree with the library type should be "impossible", while setting it to 1 says that alignments that disagree with the library type are no less likely than those that do
-g [--geneMap] arg	File containing a mapping of transcripts to genes. If this file is provided Salmon will output both quant.sf and quant.genes.sf files, where the latter contains aggregated gene-level abundance estimates. The transcript to gene mapping should be provided as either a GTF file, or a in a simple tab-delimited format where each line contains the name of a transcript and the gene to which it belongs separated by a tab. The extension of the file is used to determine how the file should be parsed. Files ending in '.gtf', '.gff' or '.gff3' are assumed to be in GTF format; files with any other extension are assumed to be in the simple format. In GTF / GFF format, the "transcript_id" is assumed to contain the transcript identifier and the "gene_id" is assumed to contain the corresponding gene identifier.
-z [--writeMappings] [=arg(=)]	If this option is provided, then the quasi-mapping results will be written out in SAM-compatible format. By default, output will be directed to stdout, but an alternative file name can be provided instead.
--meta	If you're using Salmon on a metagenomic dataset, consider setting this flag to disable parts of the abundance estimation model that make less sense for metagenomic data.

advanced options:

--alternativeInitMode	[Experimental]: Use an alternative strategy (rather than simple interpolation between) the online and uniform abundance estimates to initialize the EM / VBEM algorithm.
--disableReadMapping	The only direction of the quantification direction along which information on read-to-read alignments is used. This will be

Motivation

Almost all pieces of scientific software have tunable parameters.

- Their settings can greatly impact the quality of output.
- Default parameters are best on average but may be bad in general.

default

```
... yl-lhqflspssnqrtdqyggsvlenrarlvlevvdavcnewsad-RIGIRVSPigtfq ...
... kP-LGVKLPPyf--dlvhfdimaeilnqfpltyvsnv-nsig---nglfidpeaesv ...
... yl-lnqfldphsnttrtdeyggsvlenrarftlevvdalveaighe-KVGLRLSPygvmfn ...
... yl-plqflnpyynkrtdkyggsvlenrarfwletlekvkhavgsdcAIATRF---GVdt ...
... kvPLYVKLSPnv-tdivpiakaveaagadgltmintl-----mgvrfdlktrqp ...
```

alternate

```
... gsvenrarlvlevvdavcnewsad-RIGIRVSPigtfqnvdnngpnee--adaly1--- ...
... ydfeatekllke-----vftfftk-PLGVKLPPyf-----dlvhfdim ...
... gslenrarftlevvdalveaighe-KVGLRLSPygvmfnsmsggaetgivaqyayvage ...
... gslenrarfwletlekvkhavgsdcAIATRFGV-----dtvygpgq ...
... tdpevaaalvka-----ckavskv-PLYVKLSPnvt-----divpiaka ...
```

Motivation

Almost all pieces of scientific software have tunable parameters.

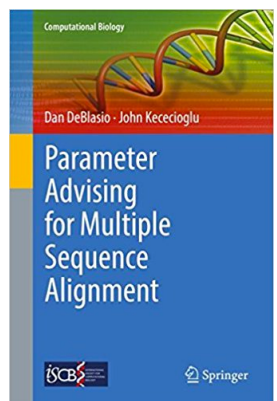
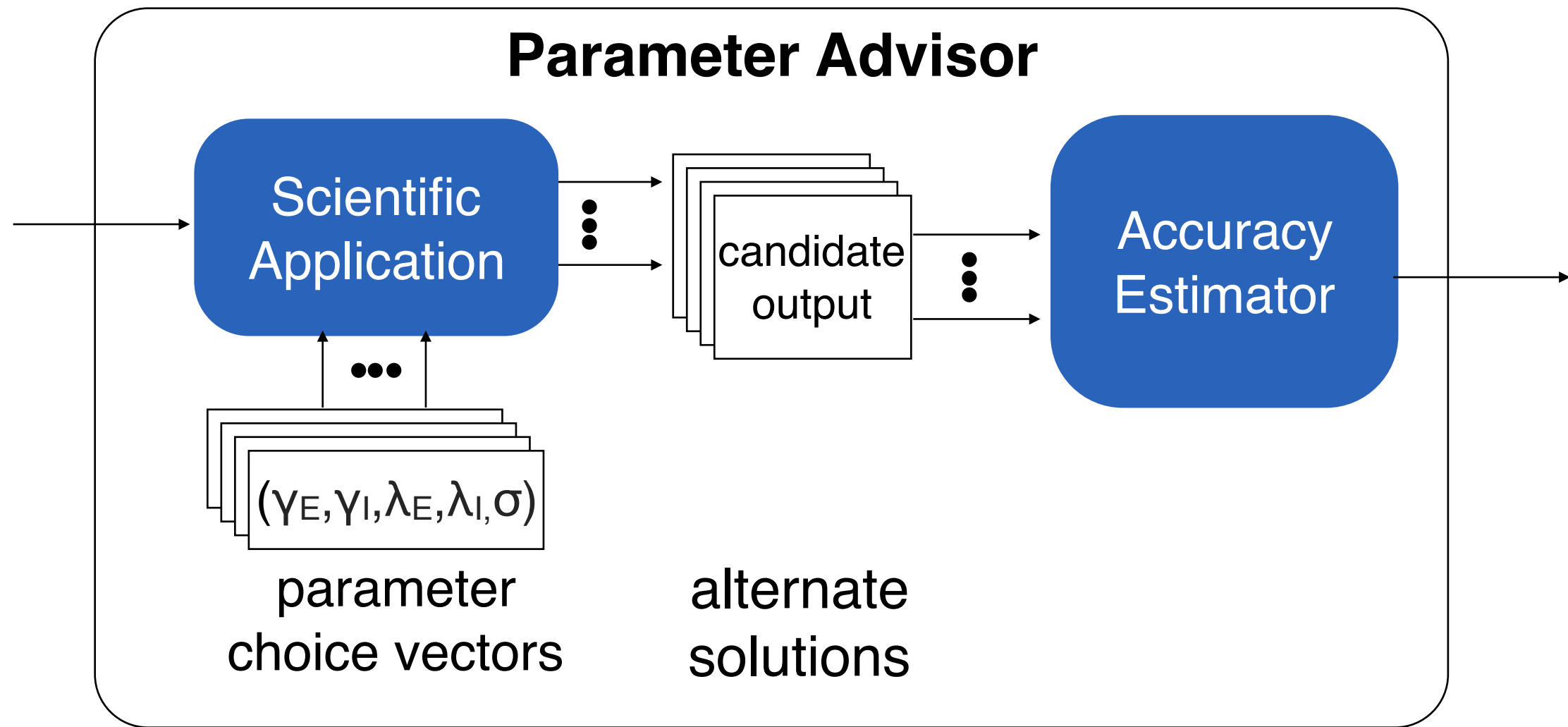
- Their settings can greatly impact the quality of output.
- Default parameters are best on average but may be bad in general.
- The most interesting problems are rarely "average".
- Mis-configuration can lead to missed or incorrect conclusions.

Can we find a parameter choice
that is best for a given input?

Parameter advising

Given an input an advisor

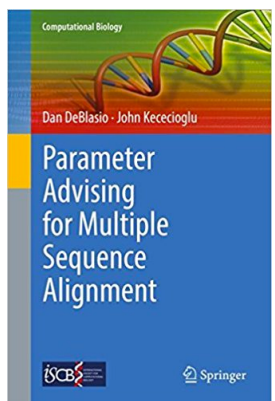
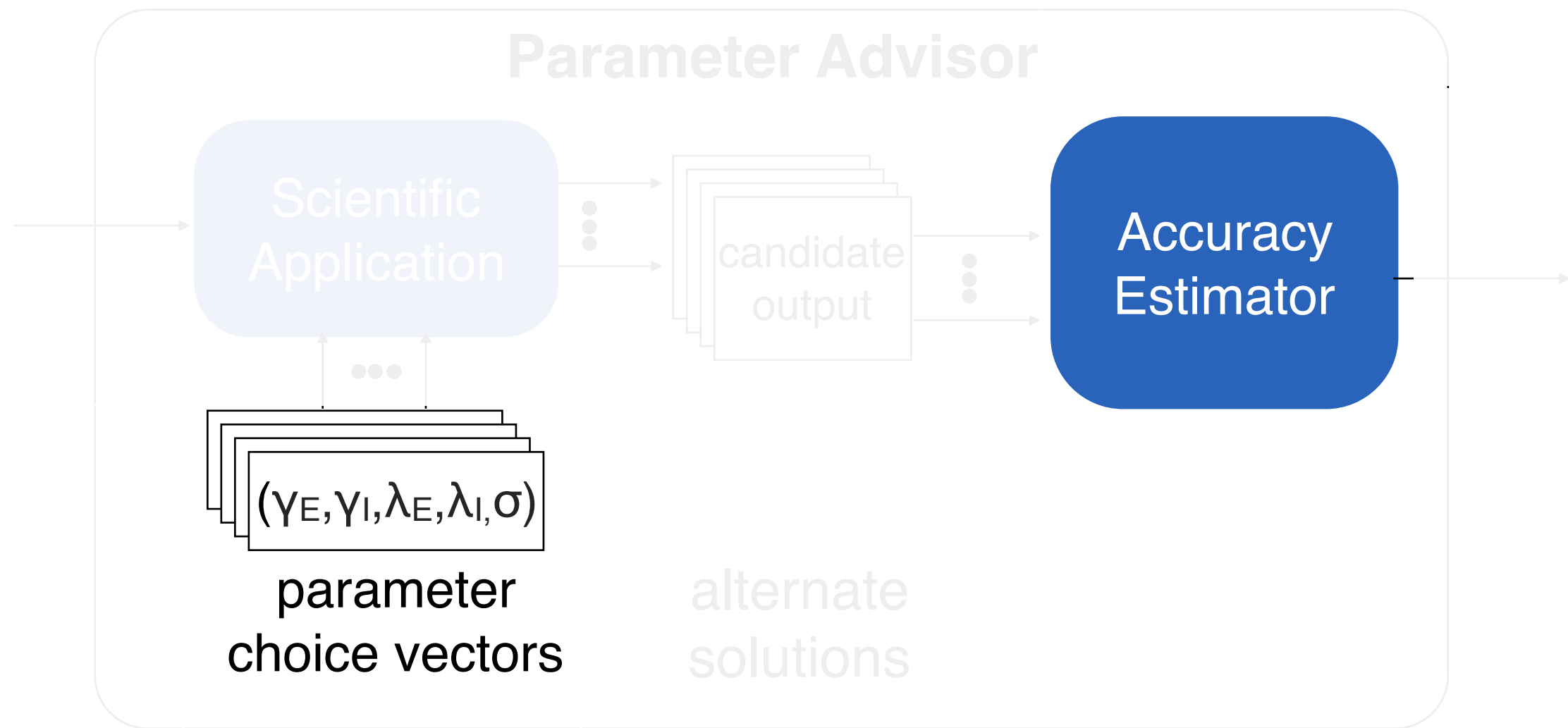
- uses an **advisor set** of parameter choice vectors to obtain candidates, then
- returns the most accurate of these solutions based on the **accuracy estimation**.



Parameter advising

Given an input an advisor

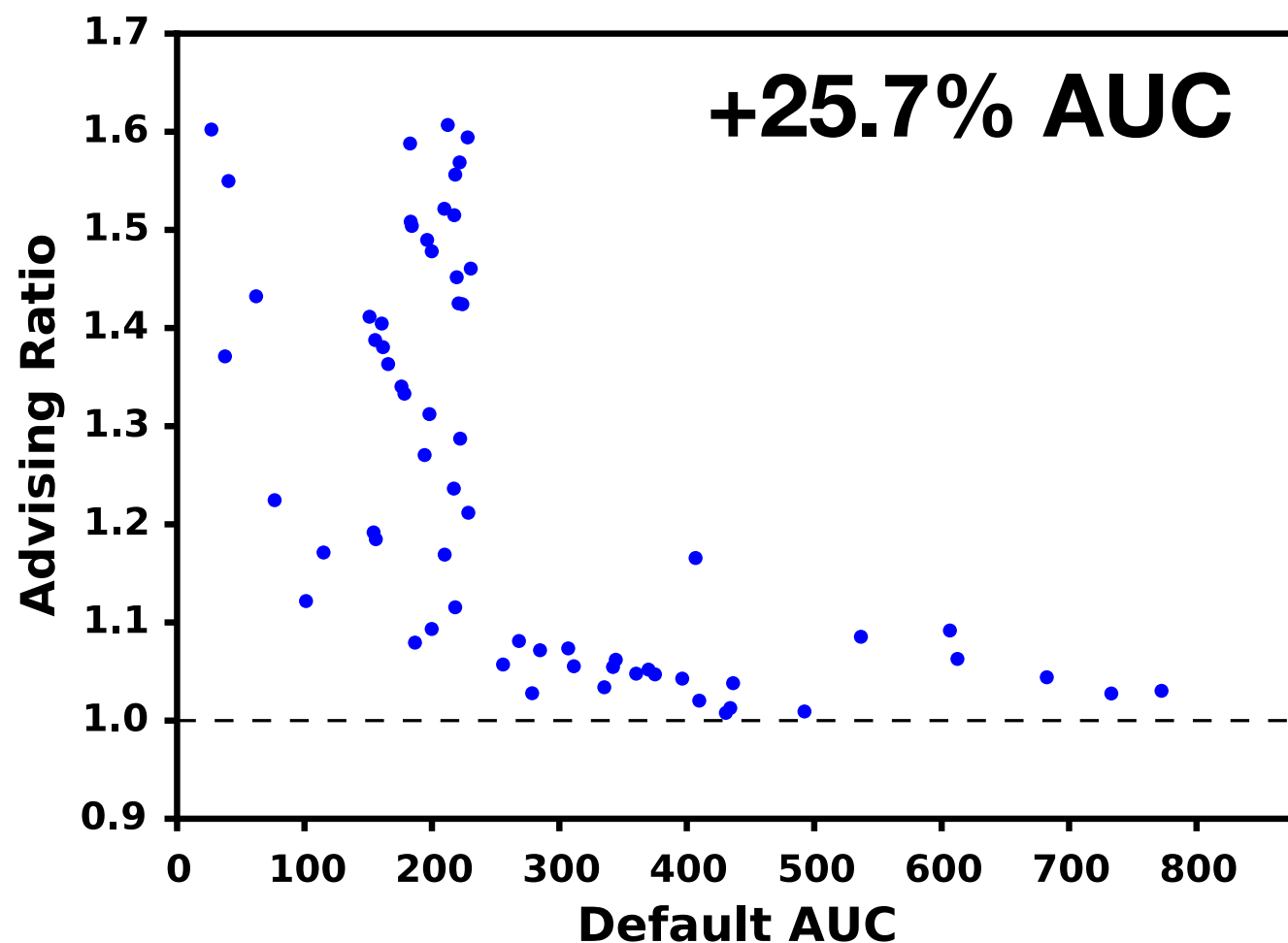
- uses an **advisor set** of parameter choice vectors to obtain candidates, then
- returns the most accurate of these solutions based on the **accuracy estimation**.



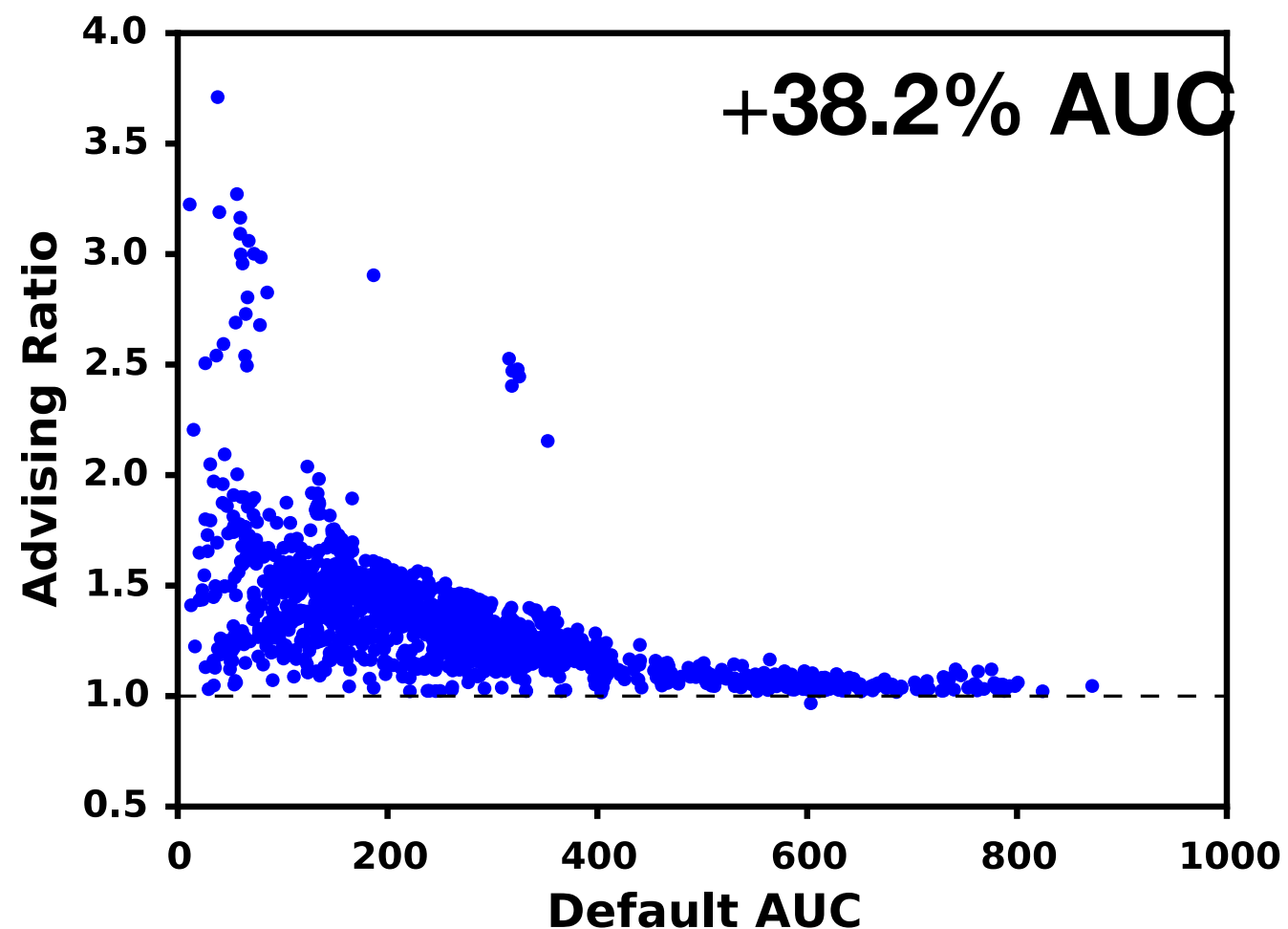
Transcript assembly advising

Reference is available, therefore Area Under the ROC Curve was used as the accuracy estimator

65 ENCODE experiments



1595 SRA experiments



Transcript assembly advising

Reference is available, therefore Area Under the ROC Curve was used as the accuracy estimator

**How can we use Parameter Advising
when a reference isn't available?**

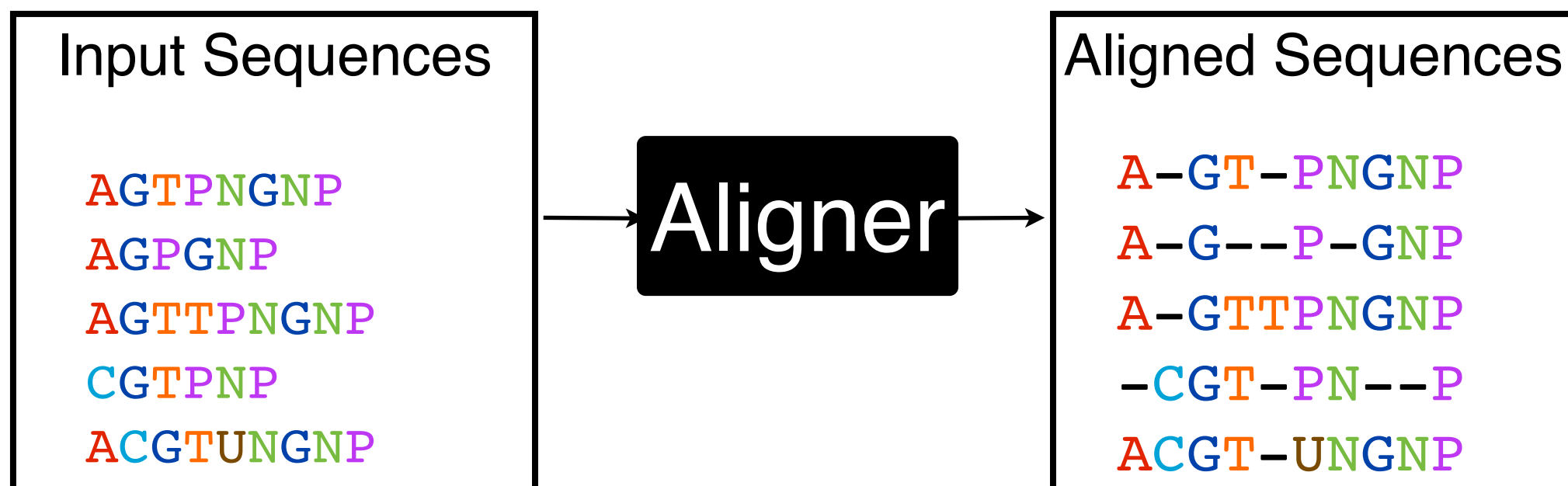
Multiple sequence alignment

Given

- a set of sequences $\{S_1, S_2, \dots, S_m\}$, and
- an alignment objective function

find an $m \times n$ matrix

- where each row represents one sequence from the set with inserted gaps, and
- is optimal under the objective function.



Motivation

Multiple sequence alignment is a **fundamental problem** in bioinformatics.

- multiple sequence alignment is **NP-Complete**
- many **popular aligners** for multiple sequence alignment
- each aligner has many **parameters** whose values affect the alignment that is output

Accuracy estimation

Alignment accuracy is measured with respect to a reference alignment.

reference alignment	computed alignment
... a D E h s a D E h - s ...
... d S R - d d S R - - d ...
... a S H l t a S - H l t ...

- accuracy is the **fraction of substitutions** from the reference that are in the computed alignment,

Accuracy estimation

Alignment accuracy is measured with respect to a reference alignment.

reference alignment	computed alignment	
... a D E h s a D E h - s ...	66% Accuracy
... d S R - d d S R - - d ...	
... a S H l t a S - H l t ...	
↑ ↑		

- accuracy is the **fraction of substitutions** from the reference that are in the computed alignment,
- measured on the **core columns** of the reference.

Accuracy estimation

Our estimator **Facet** (“**F**eature-based **AC**curacy **EsT**imator”)

- estimates accuracy by a **polynomial** on feature functions,
- efficiently learns the polynomial **coefficients** from examples,
- uses **novel features** that are efficient to evaluate.

Accuracy estimation

The estimator $E(A)$ is a **polynomial** in the feature functions $f_i(A)$.

linear estimator

$$E(A) \quad := \quad \sum_i c_i f_i(A)$$

quadratic estimator

$$E(A) \quad := \quad \sum_i c_i f_i(A) + \sum_i \sum_j c_{ij} f_i(A) f_j(A)$$

Learning the estimator

We learn the estimator using **examples** consisting of

- an **alignment**, and
- its associated **true accuracy**.

Learning the estimator

We learn the estimator using **examples** consisting of

- an **alignment**, and
- its associated **true accuracy**.

Learning finds optimal **coefficients** that either fit

- accuracy **values** of the examples, or
- accuracy **differences** on pairs of examples,
- by solving a **linear or quadratic program**.

Feature functions

We use protein alignment **feature** functions that

- are **fast** to evaluate,
- measure **novel** properties,
- use **non-local** information,
- involve **secondary structure**.

Feature functions

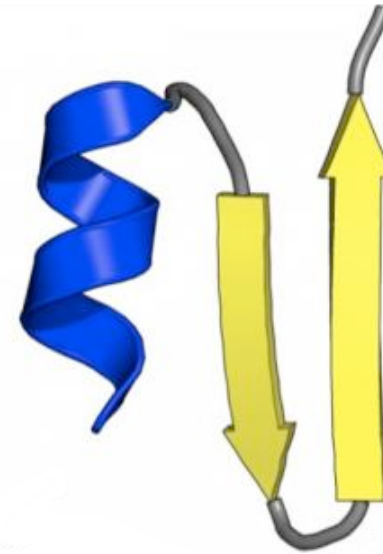
Features based **only** on the input alignment

- Amino Acid Identity
- Average Substitution Score
- Information Content
- ...

Feature functions

There are three **types** of secondary structure

- α -helix,
- β -strand,
- coil.



K	C	V	E	L	F	L	E	T	A	n	-	-	-	-	-	-	-	-	-	-	E	F	E	D	V	E	L	H	H	H	H	H	R	L	H	K	E	E	G	H	E	V	Y	I	A	S	f	e	r	g	t	i	t	g	-	-	K	h	-	g	-			
K	E	I	E	A	V	L	I	T	d	-	-	-	-	-	-	-	-	-	-	-	E	F	E	D	S	E	F	T	S	P	A	D	E	F	R	K	A	G	H	E	V	I	T	I	E	k	q	a	g	k	t	v	k	g	-	-	k	k	-	g	C			
R	E	A	L	V	E	I	L	A	k	-	-	-	-	-	-	-	-	-	-	-	G	A	E	E	E	M	E	T	V	I	P	V	D	V	M	R	R	A	G	I	K	V	T	V	A	G	l	a	g	k	d	p	v	q	C	E	-	-	s	r	-	d	-	
K	E	I	L	V	E	I	A	A	d	e	r	y	l	p	t	d	n	g	k	l	f	s	t	G	N	H	P	I	E	T	L	L	P	L	H	H	A	A	G	F	E	F	E	V	A	T	i	s	g	-	l	m	t	k	f	-	-	e	y	w	a	m		
K	E	I	L	V	E	I	A	A	d	e	r	y	l	p	t	d	n	g	k	l	f	s	t	G	N	H	P	I	E	T	L	L	P	L	H	H	A	A	G	F	E	F	E	V	A	T	i	s	g	-	l	m	t	k	f	-	-	e	y	w	a	m		
K	E	I	L	V	E	I	A	A	d	e	r	y	l	p	t	d	n	g	k	l	f	s	t	G	N	H	P	I	E	T	L	L	P	L	H	H	A	A	G	F	E	F	E	V	A	T	i	s	g	-	l	m	t	k	f	-	-	e	y	w	a	m		
K	E	V	L	L	A	L	T	s	y	n	d	v	f	y	s	-	d	g	a	-	k	t	G	V	F	V	V	E	A	L	H	H	P	F	N	T	F	R	K	E	G	F	E	V	D	F	V	S	e	t	g	-	k	-	f	g	w	-	-	d	e	h	s	l
R	E	A	L	V	E	I	L	A	k	-	-	-	-	-	-	-	-	-	-	-	-	G	A	E	E	E	M	E	T	V	I	P	V	D	V	M	R	R	A	G	I	K	V	T	V	A	G	l	a	g	k	d	p	v	q	C	E	-	-	s	r	-	d	-
K	E	I	E	G	V	E	L	S	g	-	c	g	v	y	-	-	-	-	-	-	-	d	G	S	E	I	H	E	A	V	L	T	L	L	A	I	S	R	S	G	A	Q	A	V	C	F	A	p	d	-	-	k	q	q	v	d	v	i	n	h	l	t	g	

<http://www.ebi.ac.uk/training/online/>

Feature functions

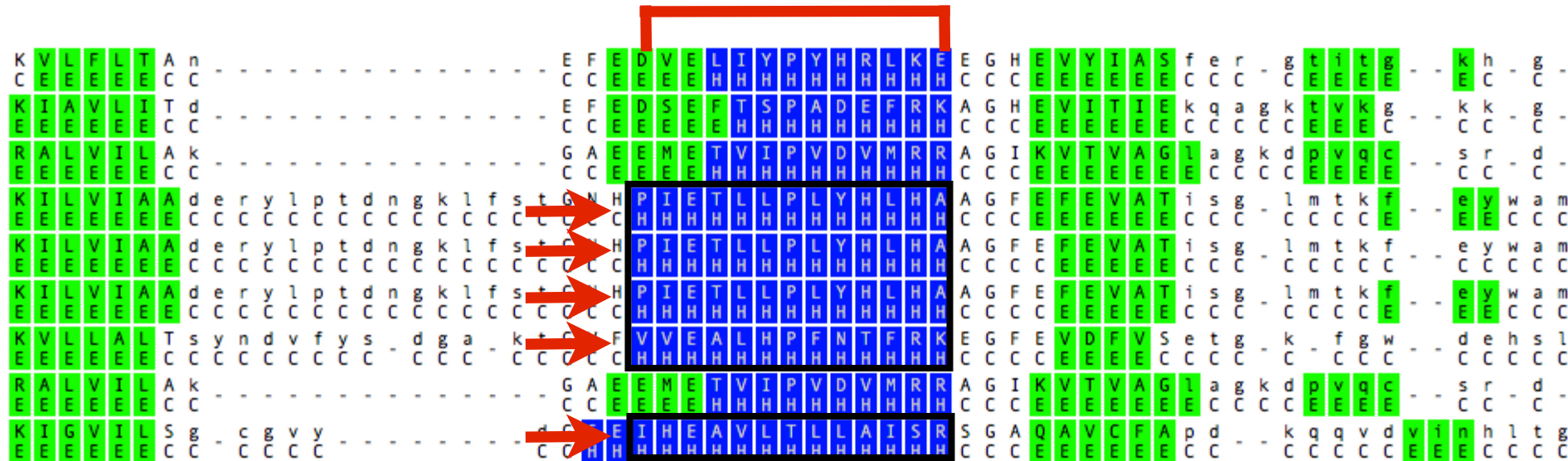
Features using predicted **secondary structure**

- Secondary Structure Percent Identity
- Secondary Structure Agreement
- Secondary Structure Blockiness
- ...

Secondary structure blockiness

A **block** B in alignment A is

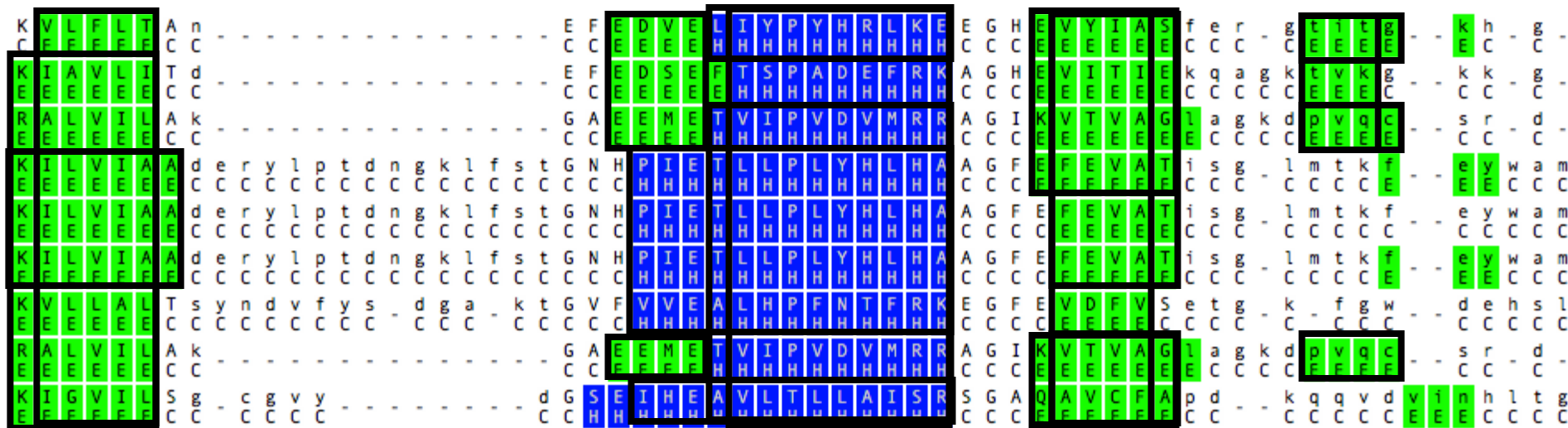
- an **interval** of at least l columns,
- a **subset** of at least k rows,
- with the **same secondary structure** for all residues in B .



Secondary structure blockiness

A **block** B in alignment A is

- an **interval** of at least l columns,
- a **subset** of at least k rows,
- with the **same secondary structure** for all residues in B .

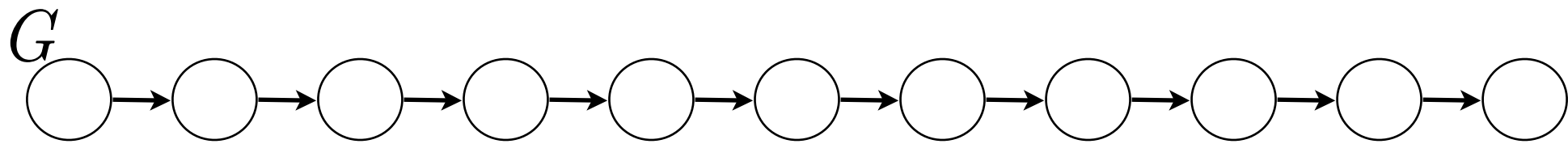


Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

Algorithm



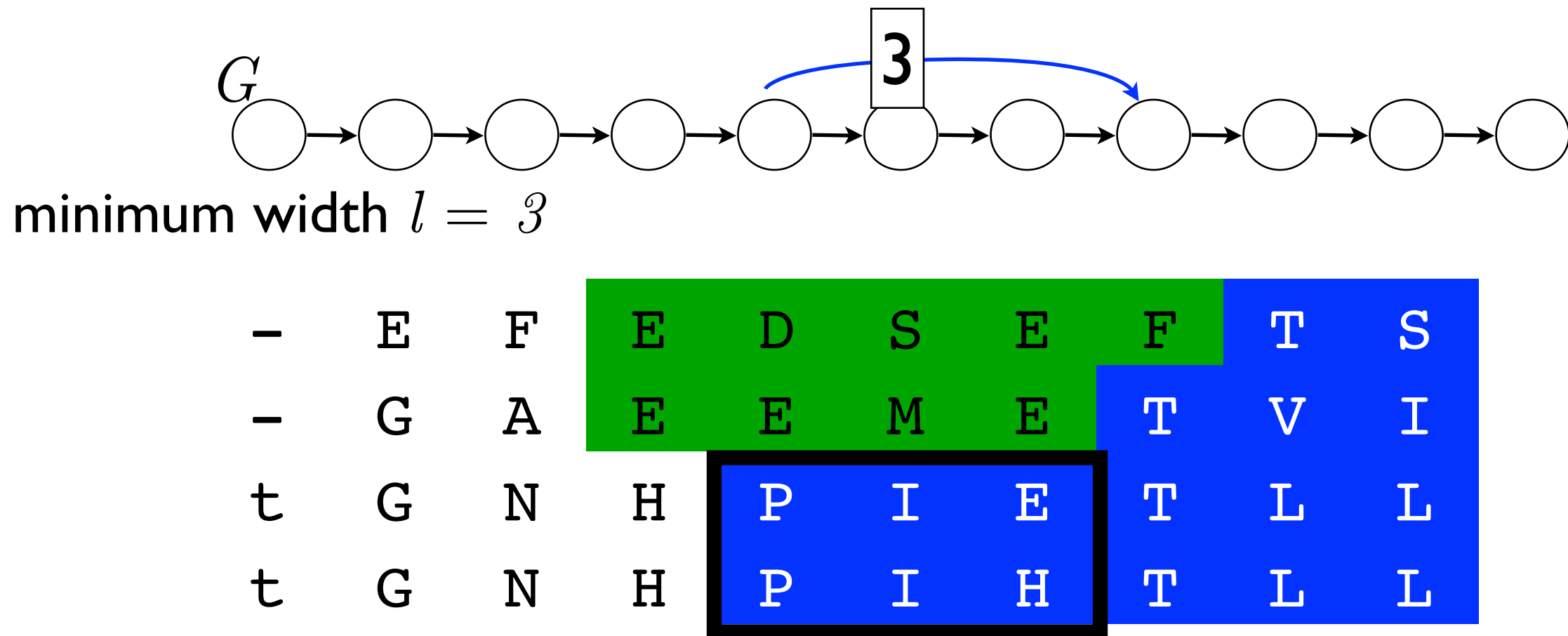
-	E	F	E	D	S	E	F	T	S
-	G	A	E	E	M	E	T	V	I
t	G	N	H	P	I	E	T	L	L
t	G	N	H	P	I	H	T	L	L

Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

Algorithm

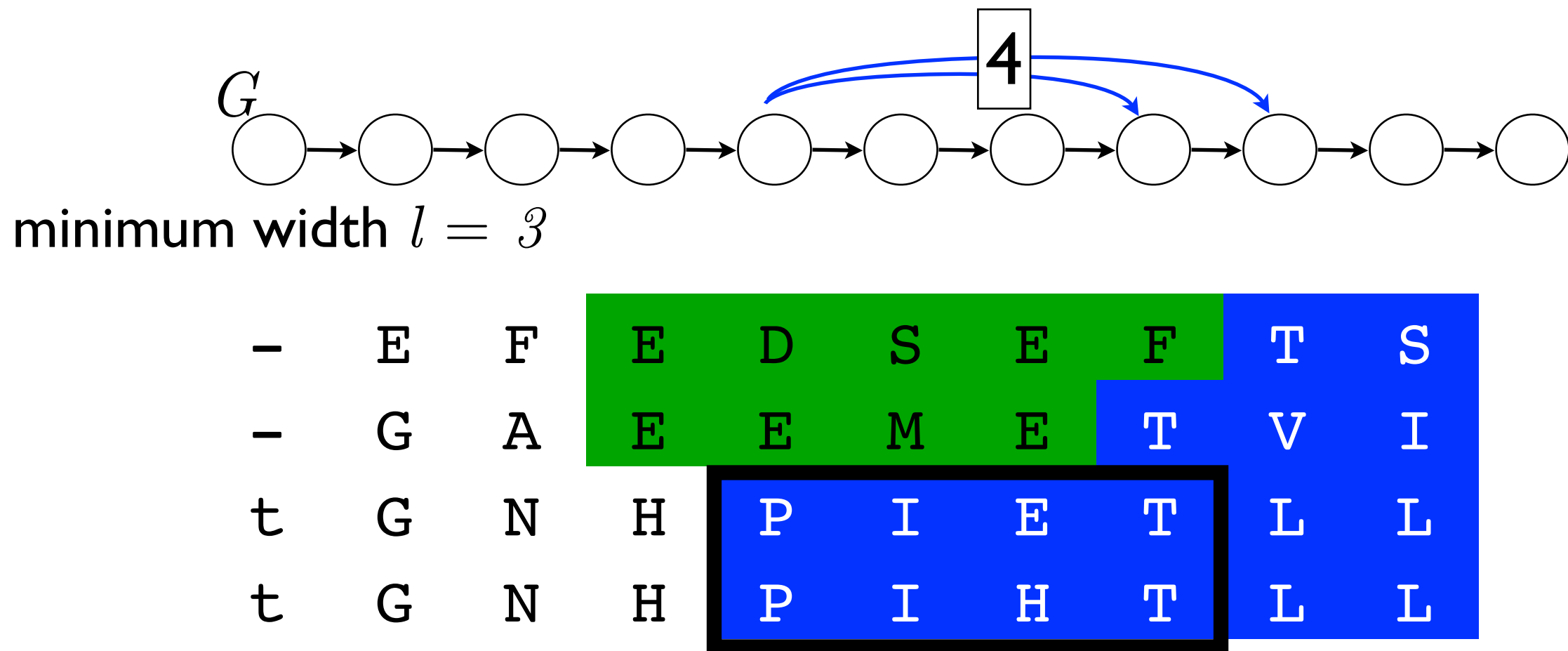


Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

Algorithm

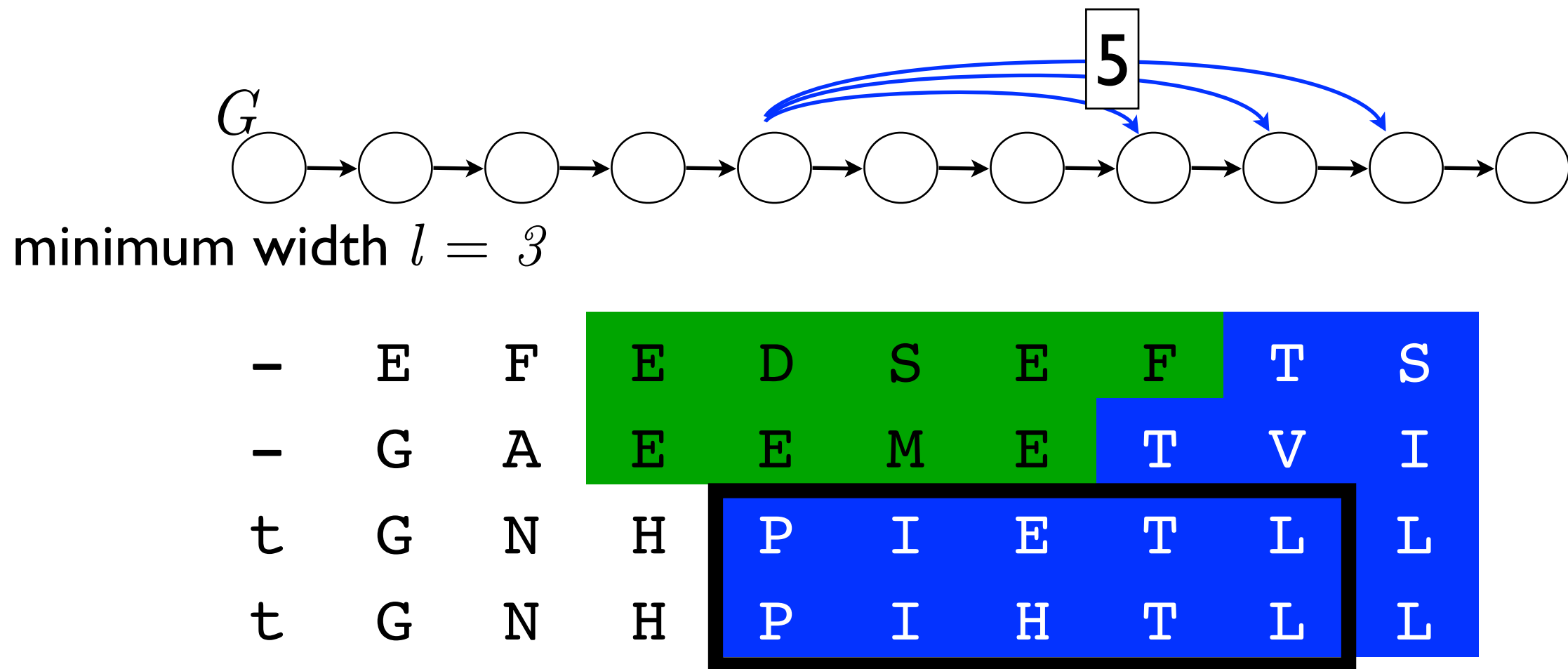


Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

Algorithm

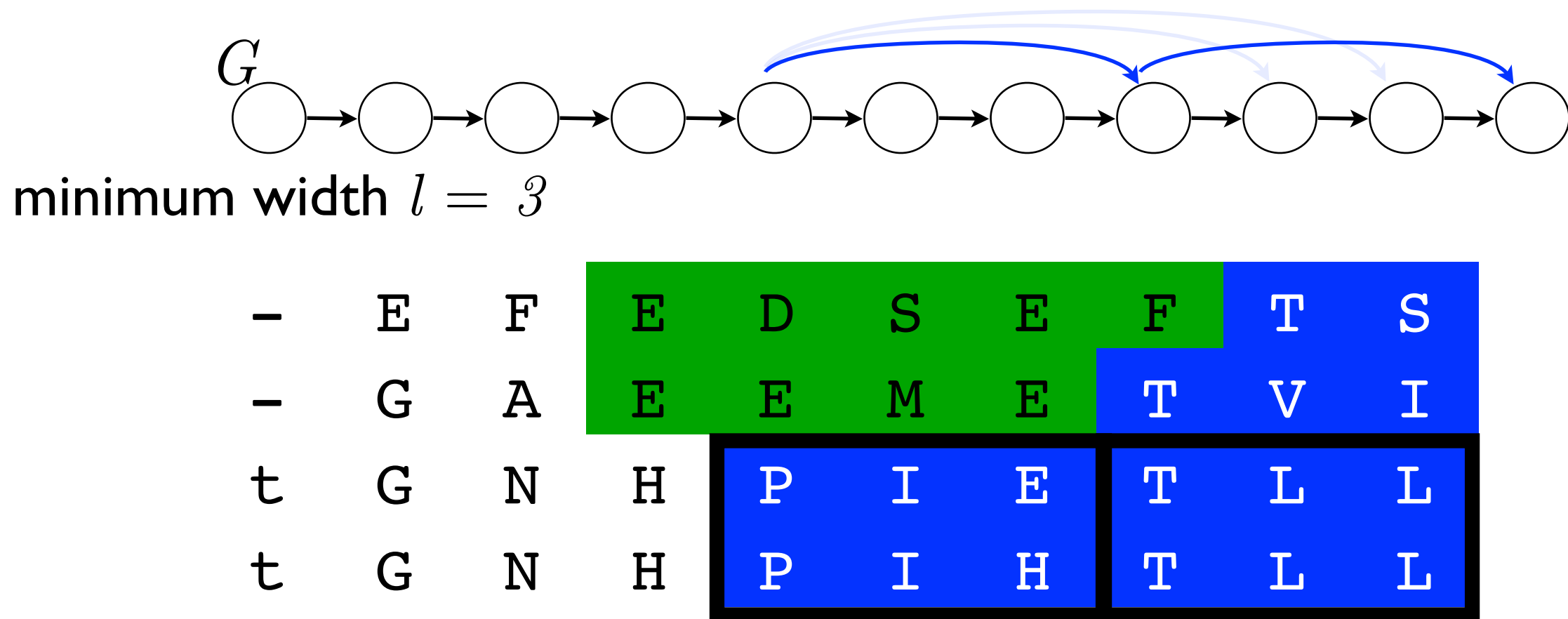


Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

Algorithm

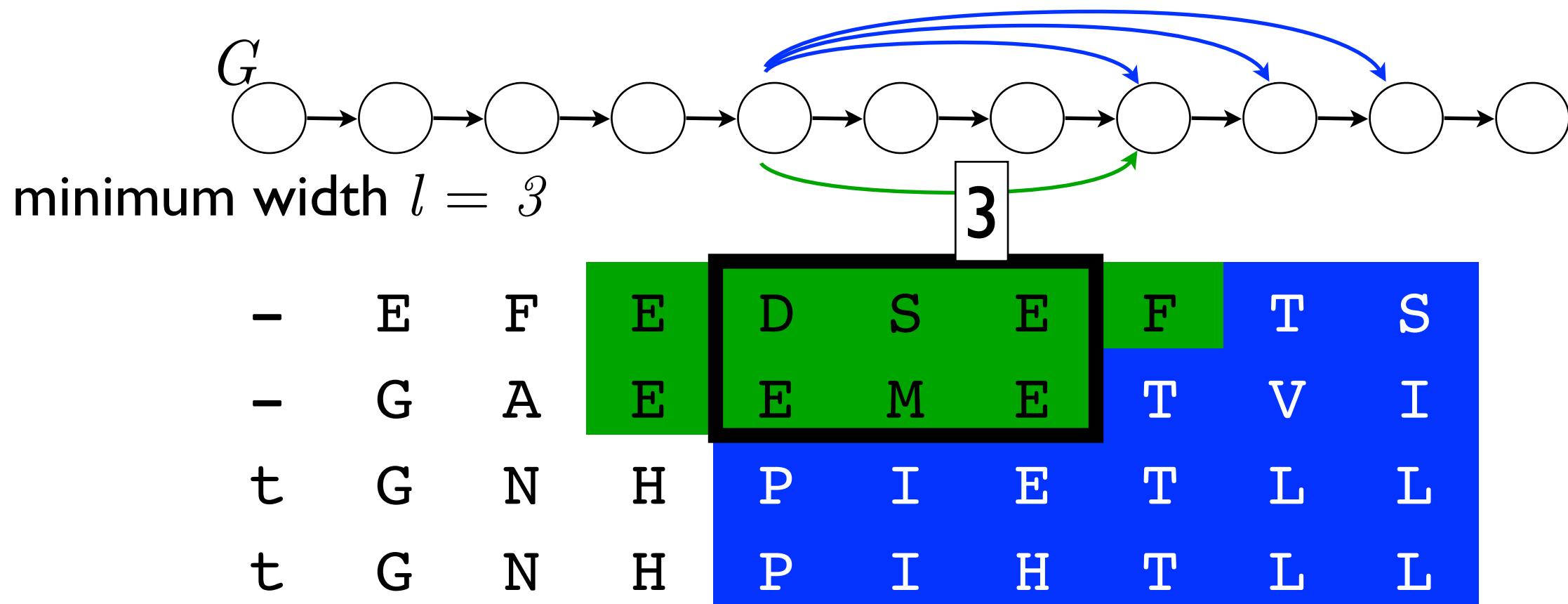


Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

Algorithm

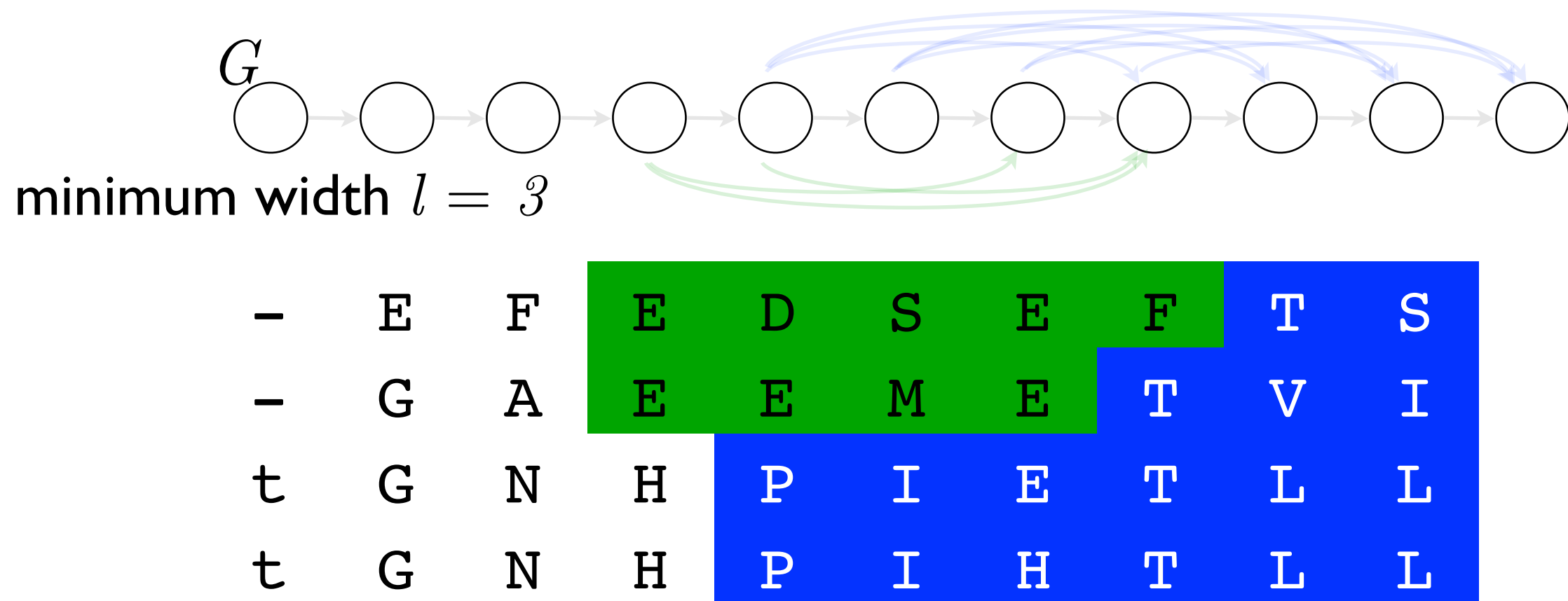


Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

Algorithm

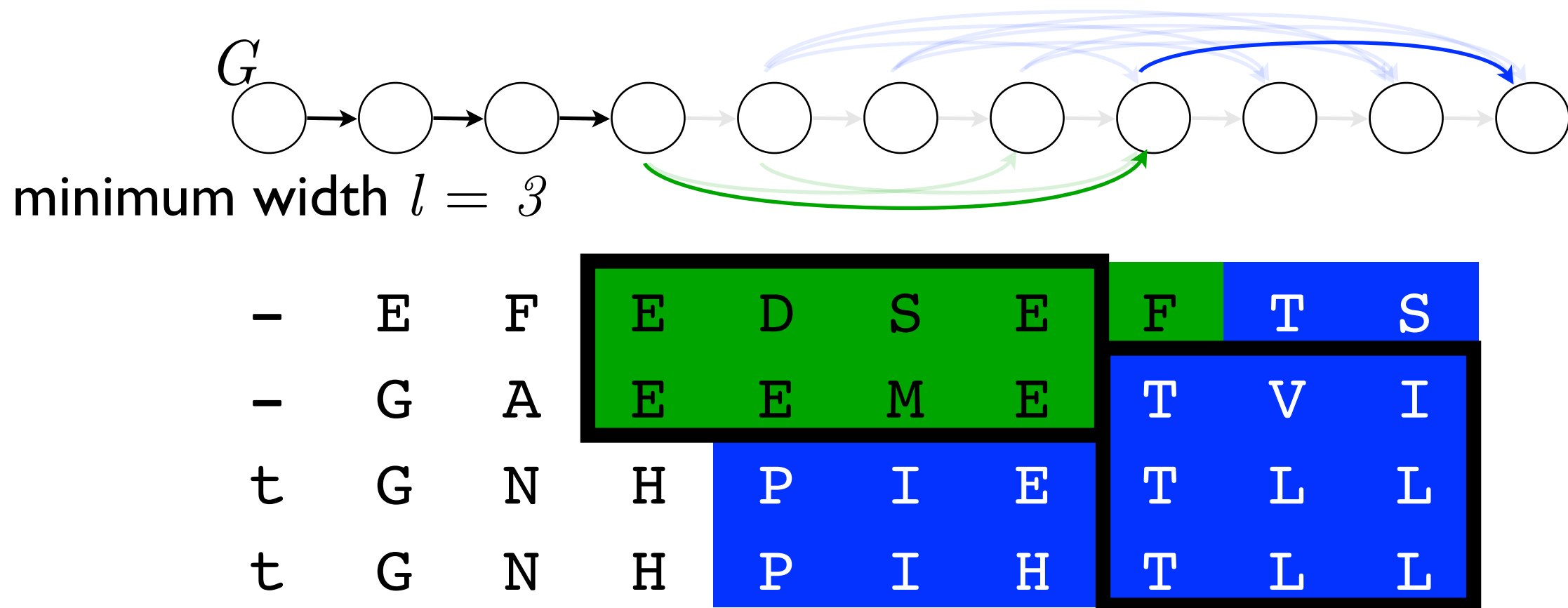


Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

Algorithm

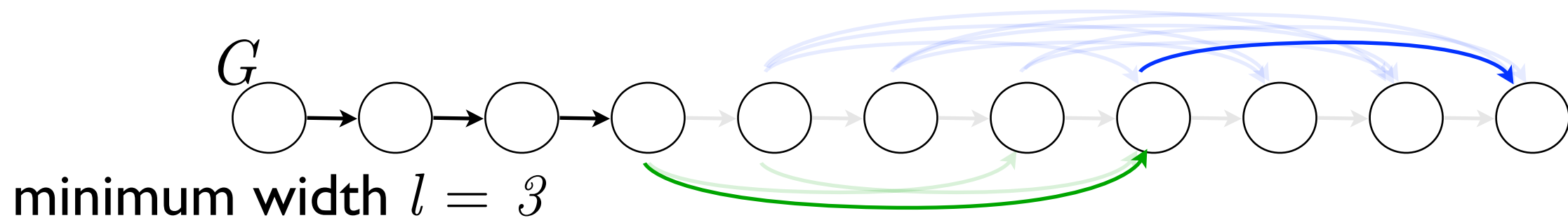


Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

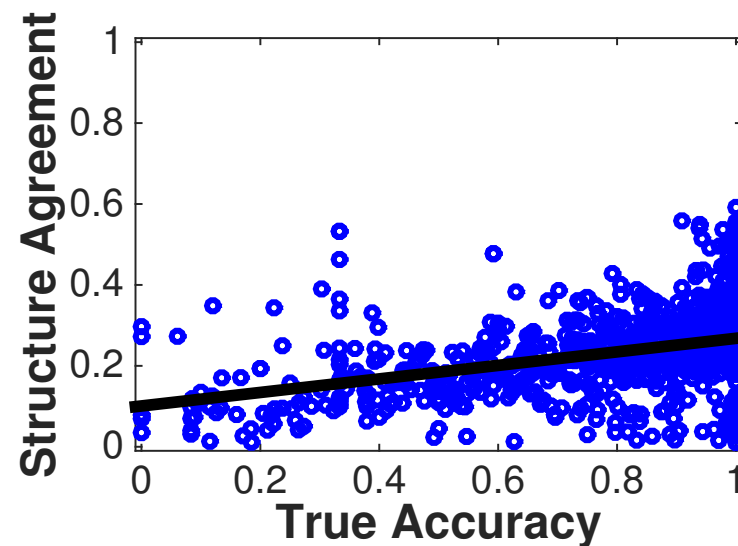
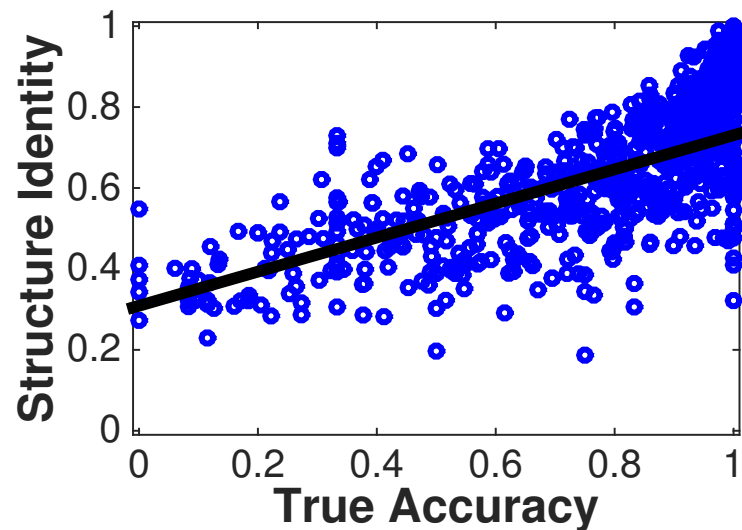
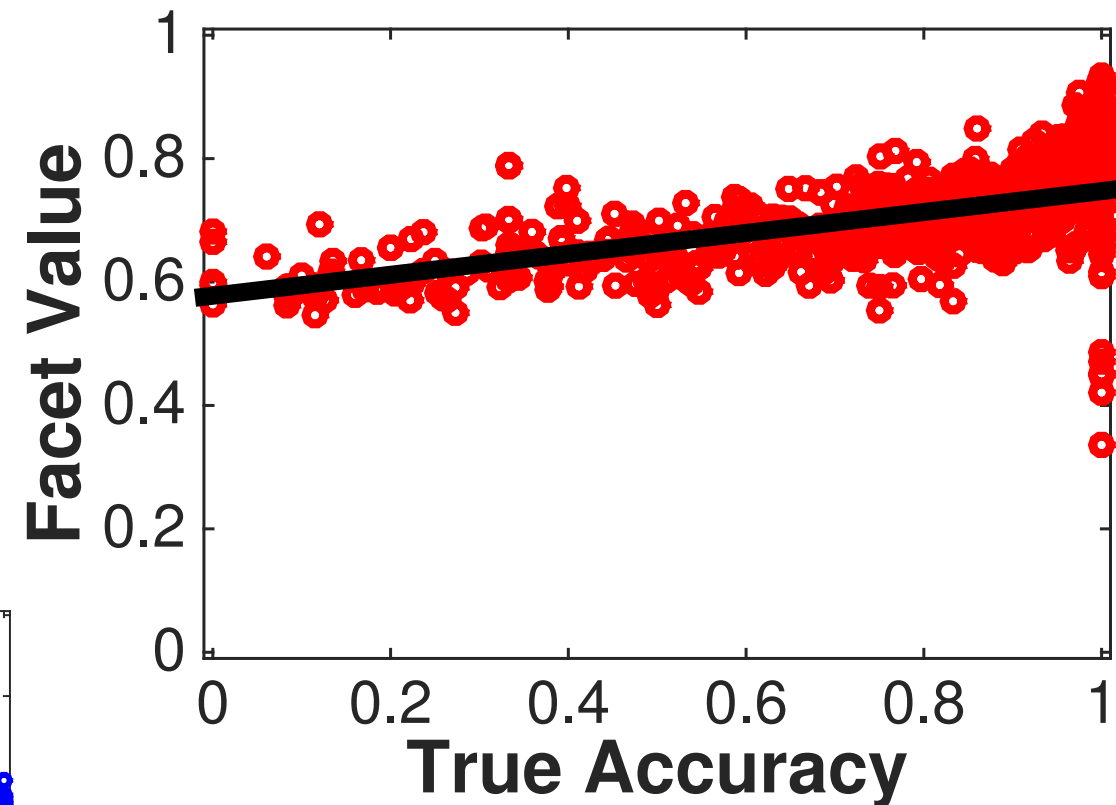
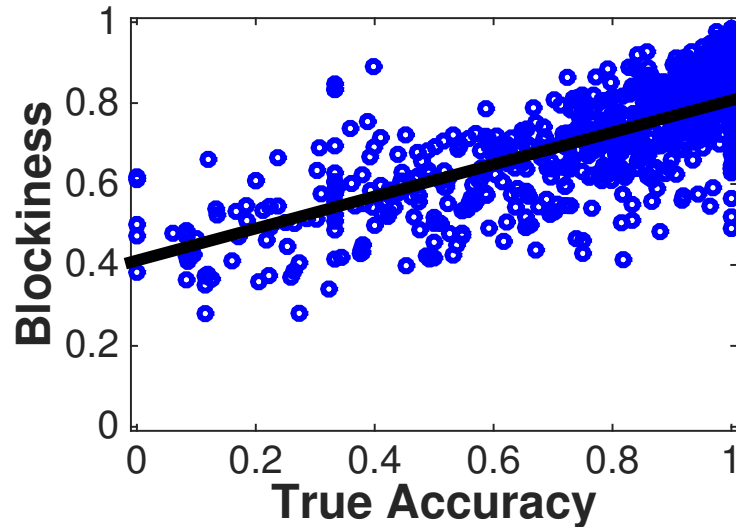
Algorithm



- Graph construction takes $O(mn)$ time.
- Graph has $O(n)$ nodes, $O(ln)$ edges
- Longest path takes $O(n)$ time.

Accuracy estimation

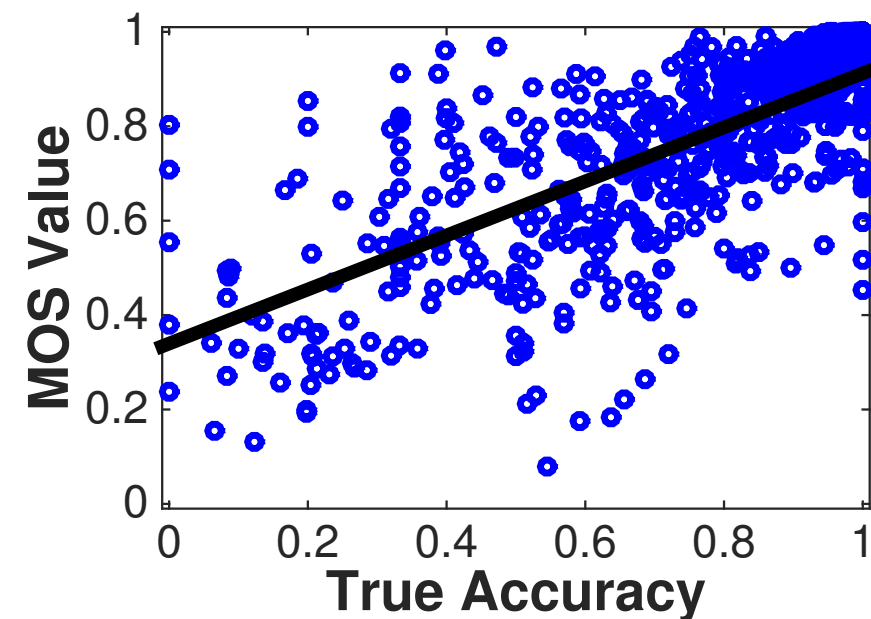
Best features trend well with accuracy.



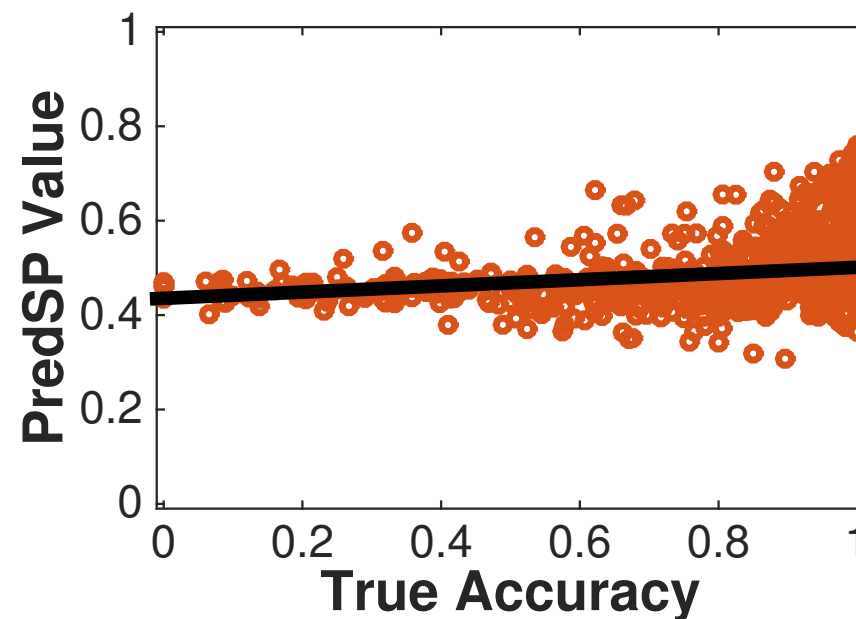
Facet estimator has less spread than its features.

Accuracy estimation

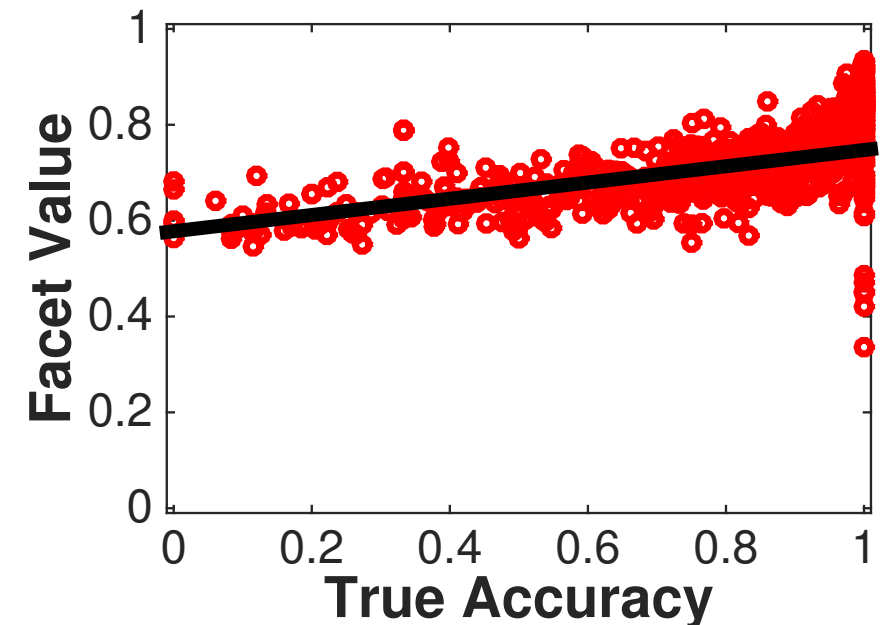
For parameter advising, an estimator should have high **slope** and low **spread**.



high slope,
high spread



low slope,
low spread



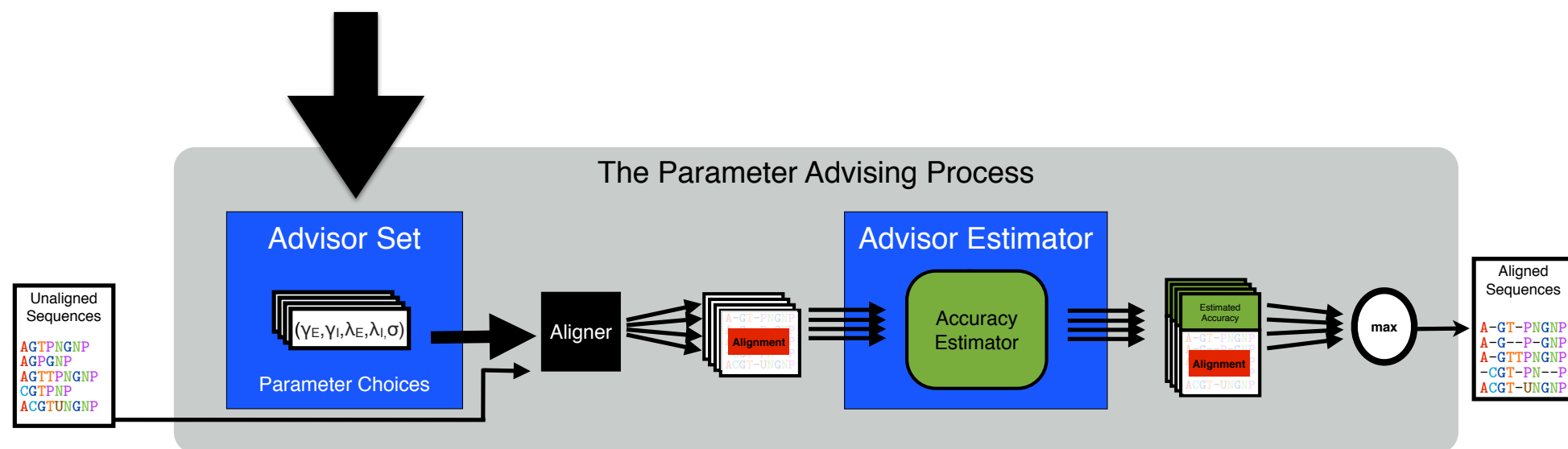
medium slope,
low spread

Facet's slope and spread is **best for advising**

Parameter advising

A **parameter advisor** has two components:

- an **accuracy estimator**, and
- a set of candidate **parameter choices**.



Given accuracy estimator,
what is the *optimal set*
of parameter choices?

Advisor Set problem

For the **Advisor Set** problem the input is

- **cardinality** bound k ,
- **universe** of parameters choices U ,
- a collection of **examples**, each with a
 - **true accuracy**,
 - **estimator value**, and
 - benchmark **weight**.

Advisor Set problem

The output is

- an optimal set $P \subseteq U$ of **parameter choices** with $|P| \leq k$, that maximizes the objective function

$$\sum_{\text{benchmarks } i} w_i \text{Accuracy}_i(P)$$

- where $\text{Accuracy}_i(P)$ is the true accuracy for benchmark i of the example chosen by an advisor that uses advisor set P

Advisor Set problem

THEOREM (Problem Complexity)

The Advisor Set problem is **NP-complete**.

- Reduction is from the **Dominating Set** problem
- **Polynomial-time** solvable for fixed k
- Optimal **oracle sets** can be found for all k in practice by integer linear programming

Approximation algorithm

THEOREM (Approximation Algorithm)

A greedy approach yields an
 $\frac{\ell}{k}$ -approximation algorithm for Advisor Set,
for constant $\ell < k$.

Approximation algorithm

THEOREM (Approximation Algorithm)

A greedy approach yields an $\frac{\ell}{k}$ -approximation algorithm for Advisor Set, for constant $\ell < k$.

The approximation ratio $\frac{\ell}{k}$ is tight.

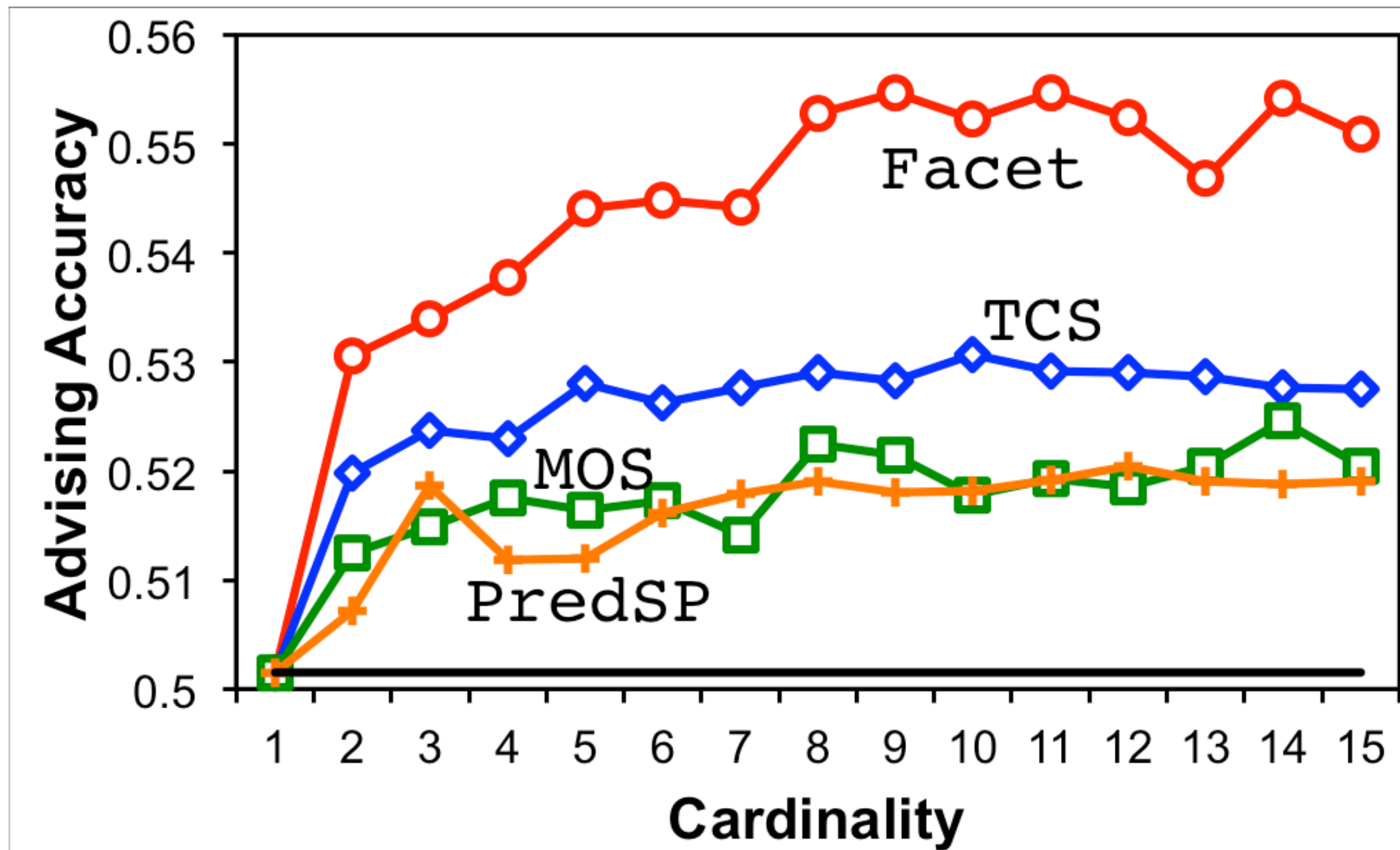
Experimental results

To **evaluate** the accuracy of advising, we consider:

- Facet and other **estimators**,
- over 800 **benchmarks**,
- evaluated with k-fold **cross validation**.

Experimental results

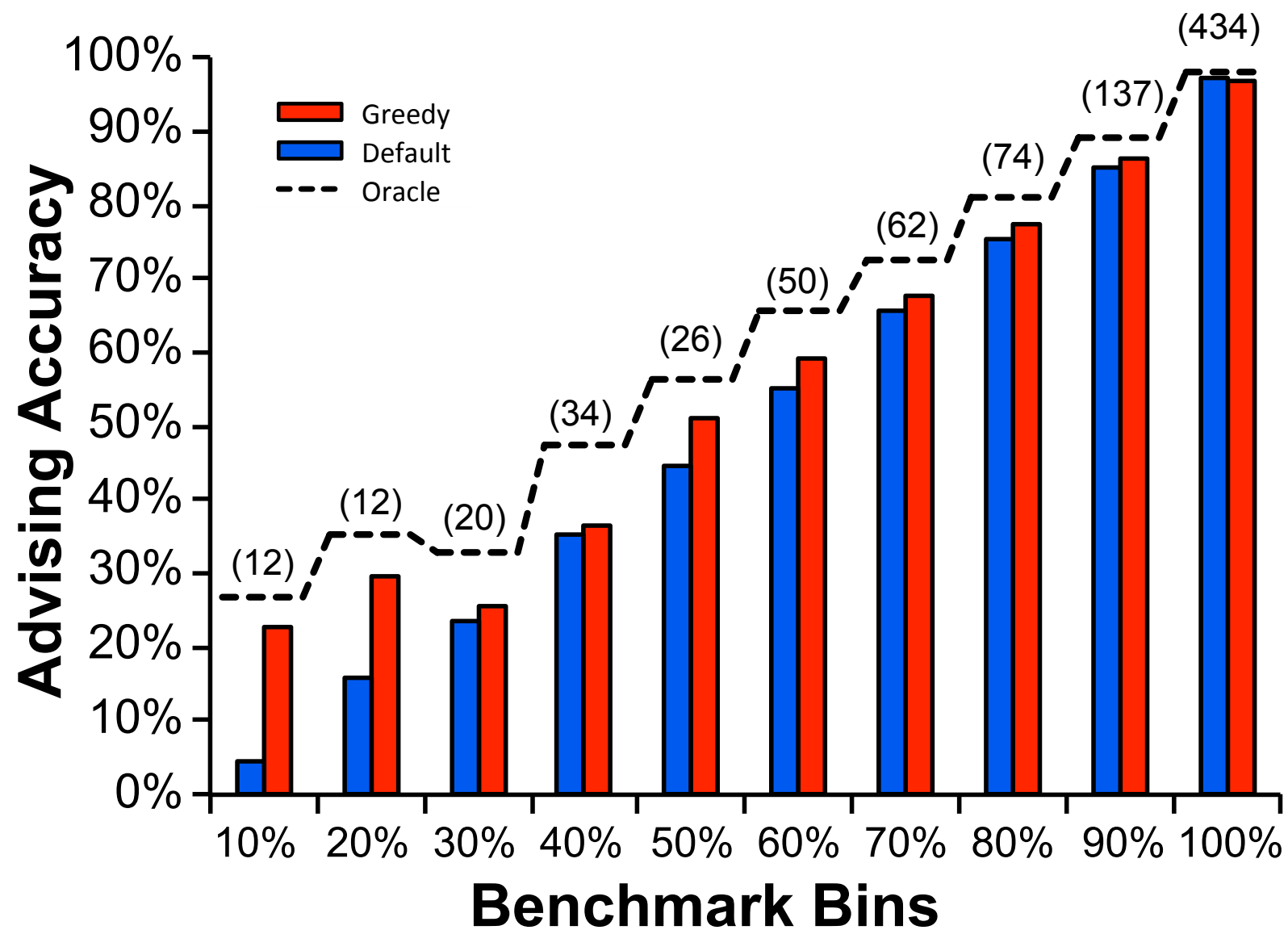
Advising performance for **various estimators**



For parameter advising Facet has the **highest accuracy**

Experimental results

Average accuracy of advisors by difficulty bin



Better
↑
greedy advising set
 $k = 10$

Boosts the accuracy on the hardest bins by almost 20%

Summary

- Parameter Advising can be used as a framework used to find input-specific parameter choices.
- Facet has the best performance for advising multiple sequence alignment.
- The Facet framework can be used to learn estimators for domains where features can be extracted.

References

Multiple Sequence Alignment Advising

- DeBlasio and Kececioglu. **Parameter Advising for Multiple Sequence Alignment.** *Springer International Publishing.* 2017.
- Kececioglu and DeBlasio. **Accuracy estimation and parameter advising for protein multiple sequence alignment.** *Journal of Computational Biology*, 20(4), April 2013.

Transcript Assembly

- DeBlasio and Kingsford. **Automatically eliminating errors induced by incorrect parameter choices for transcriptome assembly.** *bioRxiv* 2018(342865). *under review.*

Acknowledgments

Software

`facet.cs.arizona.edu`

`github.com/Kingsford-Group/scallopadvising`

Contact

`dandeblasio.com`

 danfdeblasio

Collaborators

John Kececioglu

Travis Wheeler (Montana)



Funding

US National Science Foundation

Carl Kingsford

Kwanho Kim

Jonathan King



Carnegie Mellon University
School of Computer Science

US National Science Foundation

US National Institutes of Health

Shurl and Kay Curci Foundation

Gordon and Betty Moore Foundation's
Data-Driven Discovery Initiative

Eric and Wendy Schmidt by recommendation
of the Schmidt Futures program