Building an Automated Scientist:

Three stories of accelerating scientific discovery

COMPUTER SCIENCE
University of Texas at El Paso

Dan DeBlasio
deblasiolab.org
danfdeblasio

Modern science is computational

Modern science is increasingly computational.

- Particularly in genomics, where experiments have multiple computational steps.
- Domain problems have in turn lead to algorithmic advances.

More people are relying on computational tools.

Parameter Advising for Bioinformatics

Bioinformatics software

Common themes arise in bioinformatics (and many other domain) problems.

- Many are computationally inefficient to solve exactly.
- Many tools developed for these problems.
- Each tool has many parameters whose values have an impact on the output.

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
```

```
--strictIntersect                 Modifies how orphans are assigned. When this flag is set, if the intersection of the quasi mappings for the left and right is
                                  empty, then all  mappings for the left and all mappings  for the right read are reported as  orphaned quasi-mappings
--fldMax arg                      The maximum fragment length to consider when building the empirical  distribution
--fldMean arg                     The mean used in the fragment length  distribution prior
--fldSD arg                       The standard deviation used in the  fragment length distribution prior
-f [ --forgettingFactor ] arg     The forgetting factor used in the  online learning schedule.  A smaller  value results in quicker learning, but  higher variance
                                  and may be unstable.  A larger value results in slower learning but may be more stable.  Value should  be in the interval (0.5, 1.0].
-m [ --maxOcc ] arg               (S)MEMs occuring more than this many  times won't be considered.
--initUniform                     initialize the offline inference with  uniform parameters, rather than seeding with online parameters.
-w [ --maxReadOcc ] arg           Reads "mapping" to more than this many  places won't be considered.
--noLengthCorrection              [experimental] : Entirely disables  length correction when estimating the  abundance of transcripts.  This option  can be used with
                                  protocols where one  expects that fragments derive from  their underlying targets without regard to that target's length (e.g. QuantSeq)
--noEffectiveLengthCorrection     Disables effective length correction  when computing the probability that a  fragment was generated from a  transcript.  If this flag
                                  is passed in, the fragment length distribution is not taken into account when computing this  probability.
--noFragLengthDist                [experimental] : Don't consider  concordance with the learned fragment  length distribution when trying to  determine the probability
                                  that a  fragment has originated from a  specified location.  Normally,  Fragments with unlikely lengths will be assigned a smaller
                                  relative probability than those with more likely lengths.   When this flag is passed in, the  observed fragment length has no effect
                                  on that fragment's a priori  probability.
--noBiasLengthThreshold           [experimental] : If this option is  enabled, then no (lower) threshold will be set on how short bias correction can make effective
                                  lengths. This can  increase the precision of bias  correction, but harm robustness.  The  default correction applies a threshold.
--numBiasSamples arg              Number of fragment mappings to use when learning the sequence-specific bias  model.
--numAuxModelSamples arg          The first <numAuxModelSamples> are used to train the auxiliary model parameters (e.g. fragment length distribution,  bias, etc.).
                                  After ther first  <numAuxModelSamples> observations the  auxiliary model parameters will be  assumed to have converged and will be
                                  fixed.
--numPreAuxModelSamples arg       The first <numPreAuxModelSamples> will  have their assignment likelihoods and  contributions to the transcript  abundances computed
                                  without applying  any auxiliary models.  The purpose of  ignoring the auxiliary models for the  first <numPreAuxModelSamples>  observations
                                  is to avoid applying these models before thier parameters have  been learned sufficiently well.
--useVBOpt                        Use the Variational Bayesian EM rather  than the traditional EM algorithm for  optimization in the batch passes.
--rangeFactorizationBins arg      Factorizes the likelihood used in  quantification by adopting a new notion of equivalence classes based on the  conditional probabilities
                                  with which  fragments are generated from different  transcripts.  This is a more  fine-grained factorization than the  normal rich
                                  equivalence classes.  The  default value (0) corresponds to the  standard rich equivalence classes, and  larger values imply a more
                                  fine-grained factorization.  If range factorization  is enabled, a common value to select  for this parameter is 4.
--numGibbsSamples arg             Number of Gibbs sampling rounds to  perform.
--numBootstraps arg               Number of bootstrap samples to  generate. Note: This is mutually  exclusive with Gibbs sampling.
--thinningFactor arg              Number of steps to discard for every  sample kept from the Gibbs chain. The  larger this number, the less chance  that subsequent samples
                                  are  auto-correlated, but the slower  sampling becomes.
-q [ --quiet ]                    Be quiet while doing quantification  (don't write informative output to the  console unless something goes wrong).
--perTranscriptPrior              The prior (either the default or the  argument provided via --vbPrior) will  be interpreted as a transcript-level  prior (i.e. each
                                  transcript will be  given a prior read count of this value)
--vbPrior arg                     The prior that will be used in the VBEM algorithm.  This is interpreted as a  per-nucleotide prior, unless the  --perTranscriptPrior flag
                                  is also  given, in which case this is used as a  transcript-level prior
--writeOrphanLinks                Write the transcripts that are linked  by orphaned reads.
--writeUnmappedNames              Write the names of un-mapped reads to  the file unmapped_names.txt in the  auxiliary directory.
-x [ --quasiCoverage ] arg        [Experimental]: The fraction of the  read that must be covered by MMPs (of  length >= 31) if this read is to be  considered as "mapped".
                                  This may help  to avoid "spurious" mappings. A value  of 0 (the default) denotes no coverage  threshold (a single 31-mer can yield a  mapping).
                                  Since coverage by exact  matching, large, MMPs is a rather  strict condition, this value should  likely be set to something low, if  used.
```

Tunable parameters

Tunable parameters

Biostars — BIOINFORMATICS EXPLAINED

Welcome to Biostar! Community Log In Sign Up

Question: Salmon libtype - Does it matter?

1

Hi,

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

Thanks, Matt

salmon alignment

ADD COMMENT • link •

Question: (Closed) Questions about Salmon

0

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

1. If one will probably map the reads in order to visuali
alignment based mapping so that Salmon agrees w
2. Our experiments involve looking at ribosome profilin
read counts. We also have the corresponding RNA-
one/both data sets? My idea was to not use either c
ribosome profiling data set.
3. Could someone provide a better explanation for fldM
changing the default settings be more useful? I had

Question: Salmon unmapped reads

Hi All

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

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

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

12 weeks ago by Sharon • 150

apping, but couldn't

Question: RNA-Seq analysis using STAR and Salmon

Hello! I am having some trouble figuring out how to use Salmon. I have around 30 different samples which I trimmed using bbmap then aligned them using STAR. I have all of the BAM files from this alignment. Should I

so
I am
, that

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 diffe
(promastigotes and amastigotes). For that I downloaded two RNA-seq data from SRA and
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 t
value i.e. 0,413*reads length or if there's any other advisable value. I heard that some peo*
length whereas I also found this post where is is metioned between 0.5 and 0.66*reads le

nths ag

Question: Salmon very low mapping

Hi All

low mapping rate ? #160

Open atasub opened this issue on Oct 6, 2017 · 7 comments

atasub commented on Oct 6, 2017 • edited by rob-p

I recently ran Salmon by quasi-mapping-based mode and when I checked the salmon_quant.log file, saw that mapping rate was around ~%65-68 for all of the samples. Do you have any suggestions to improve the mapping rate? I used "--libType A" to to infer the library type info and got a warning that "Greater than 5% of the fragments disagreed with the provided library typ", but I guess this is not an issue. This is an example for one of the "lib_format_counts.json" files:

Tunable parameters

Most users rely on the default parameter settings,

- which are meant to work well on average,
- but the most interesting examples are not typically "average".

Tunable parameters

Most users rely on the default parameter settings,

- which are meant to work well on average,
- but the most interesting examples are not typically "average".

default

```
... yl-lhqflspssnqrtdqyggsvenrarlvlevvdavcnewsad-RIGIRVSPigtfq ...
... kP-LGVKLPPyf--dlvhfdimaeilnqfpltyvsnv-nsig----nglfidpeaesv ...
... yl-lnqfldphsntrtdeyggsienrarftlevvdalveaighe-KVGLRLSPygvfn ...
... yl-plqflnpyynkrtdkyggslenrarfwletlekvkhavgsdcAIATRF---GVdt ...
... kvPLYVKLSPnv-tdivpiakaveaagadgltmintl---------mgvrfdlktrqp ...
```

alternate

```
... gsvenrarlvlevvdavcnewsad-RIGIRVSPigtfqnvdngpnee--adalyl--- ...
... ydfeatekllke-----vftfftk-PLGVKLPPyf---------------dlvhfdim ...
... gsienrarftlevvdalveaighe-KVGLRLSPygvfnsmsggaetgivaqyayvage ...
... gslenrarfwletlekvkhavgsdcAIATRFGV-----------------dtvygpgq ...
... tdpevaaalvka-----ckavskv-PLYVKLSPnvt--------------divpiaka ...
```

The default parameter choices misaligns this region of the sequences.

Tunable parameters

It's not just a problem in computational biology!

SATzilla: Portfolio-based Algorithm Selection for SAT

Lin Xu — XULIN730@CS.UBC.CA
Frank Hutter — HUTTER@CS.UBC.CA
Holger H. Hoos — HOOS@CS.UBC.CA
Kevin Leyton-Brown — KEVINLB@CS.UBC.CA
Department of Computer Science
University of British Columbia
201-2366 Main Mall, BC V6T 1Z4, CANADA

Journal of Artificial Intelligence Research 36 (2009) 267-306 — Submitted 06/09; published 10/09

ParamILS: An Automatic Algorithm Configuration Framework

Frank Hutter — HUTTER@CS.UBC.CA
Holger H. Hoos — HOOS@CS.UBC.CA
Kevin Leyton-Brown — KEVINLB@CS.UBC.CA
University of British Columbia, 2366 Main Mall
Vancouver, BC, V6T1Z4, Canada

Thomas Stützle — STUETZLE@ULB.AC.BE
Université Libre de Bruxelles, CoDE, IRIDIA
Av. F. Roosevelt 50 B-1050 Brussels, Belgium

Concertio — Home — Solutions — Blog

Concertio Launches Optimizer Studio to Help Performance Engineers and IT Professionals Achieve Peak System Performance

by admin | Feb 22, 2018 | News | 0 comments

Swarm and Evolutionary Computation 1 (2011) 19–31

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

journal homepage: www.elsevier.com/locate/swevo

Invited paper

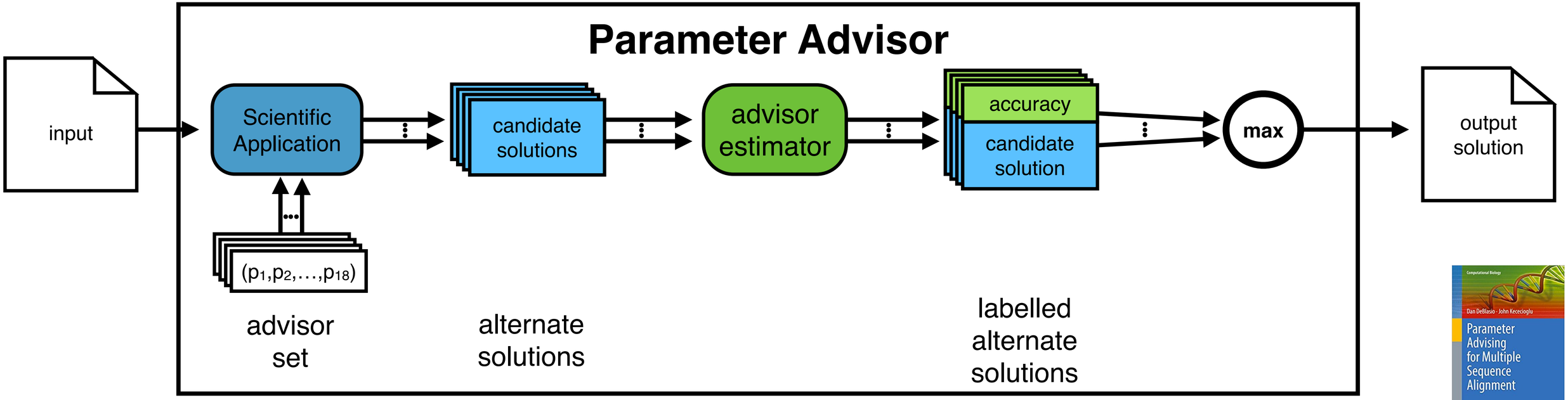
Parameter tuning for configuring and analyzing evolutionary algorithms

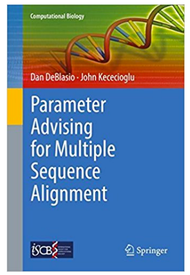
A.E. Eiben *, S.K. Smit [1]
Department of Computer Science, Vrije Universiteit Amsterdam De Boelelaan 1081a 1081 HV, Amsterdam, Netherlands

ARTICLE INFO — ABSTRACT

Parameter advising framework

Steps of advising:



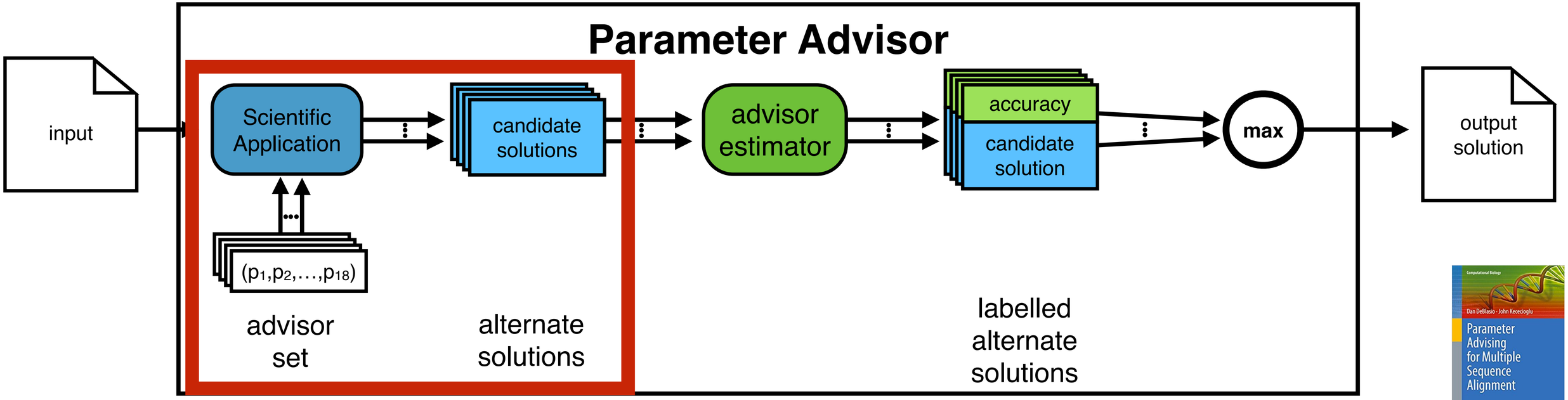


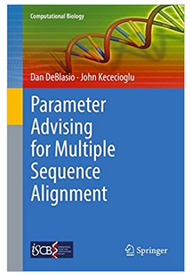
[DK, Springer International, 2017]

Parameter advising framework

Steps of advising:

- An advisor set of parameter choice vectors is used to obtain candidates.



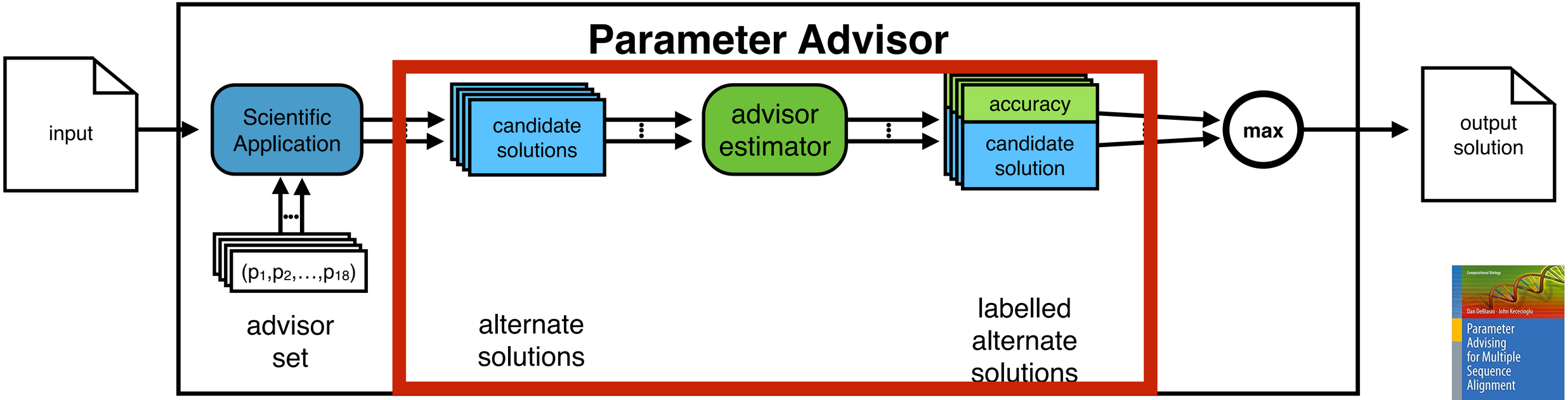


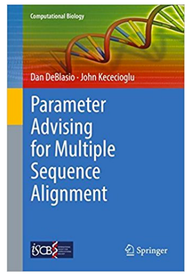
[DK, Springer International, 2017]

Parameter advising framework

Steps of advising:

- An advisor set of parameter choice vectors is used to obtain candidates.
- Solutions are ranked based on the accuracy estimation.



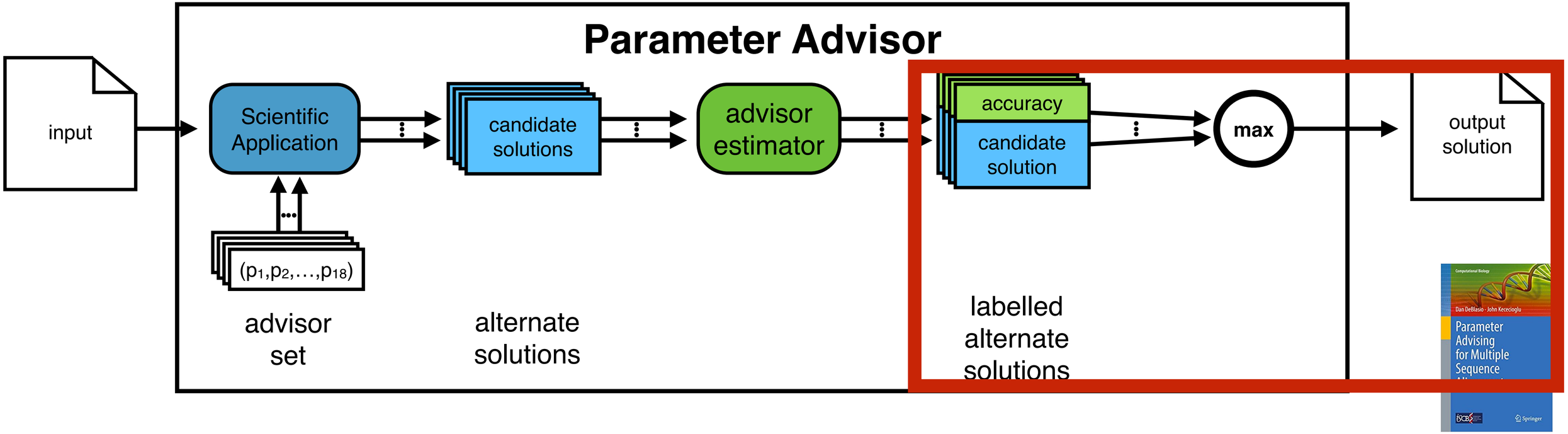


[DK, Springer International, 2017]

Parameter advising framework

Steps of advising:

- An advisor set of parameter choice vectors is used to obtain candidates.
- Solutions are ranked based on the accuracy estimation.
- The highest ranked candidate is returned.

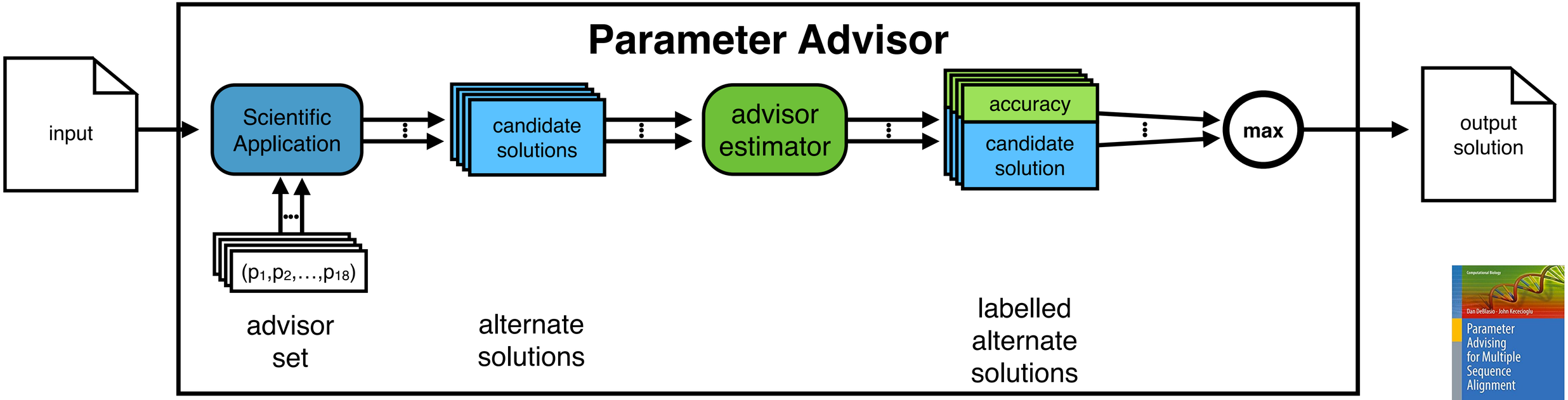


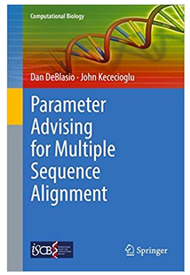
[DK, Springer International, 2017]

Parameter advising framework

Steps of advising:

- An advisor set of parameter choice vectors is used to obtain candidates.
- Solutions are ranked based on the accuracy estimation.
- The highest ranked candidate is returned.



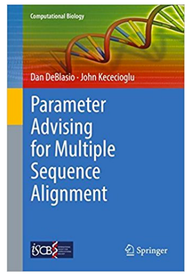


[DK, Springer International, 2017]

Parameter advising framework

Steps of advising:

- An advisor set of parameter choice vectors is used to obtain candidates.
- Solutions are ranked based on the accuracy estimation.
- The highest ranked candidate is returned.

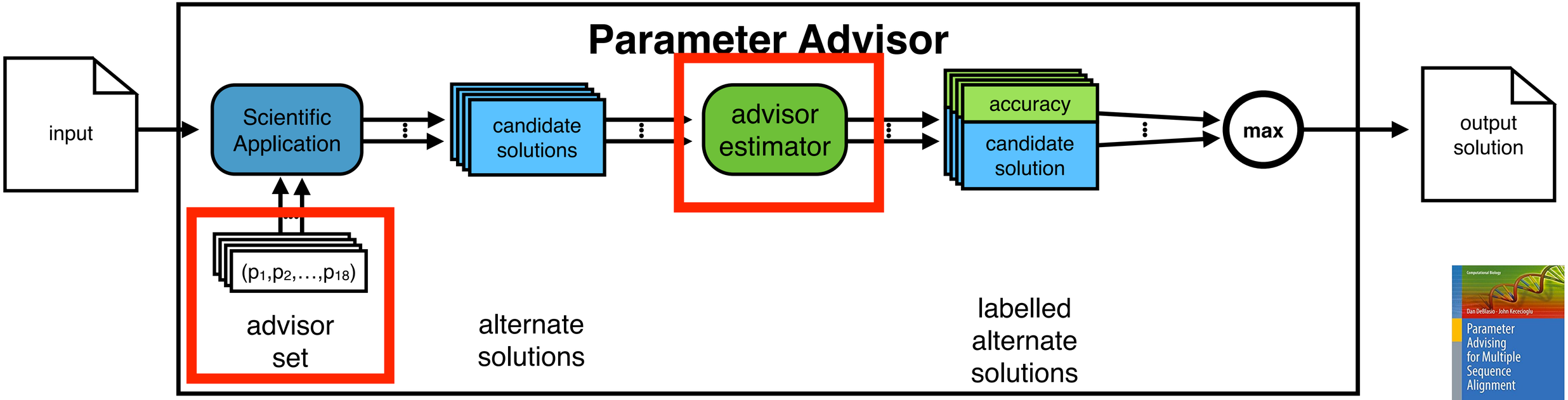


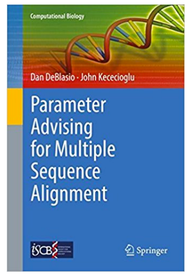
[DK, Springer International, 2017]

Parameter advising framework

Components of an advisor:

- An advisor set of parameter choice vectors.
- An advisor estimator to rank solutions.





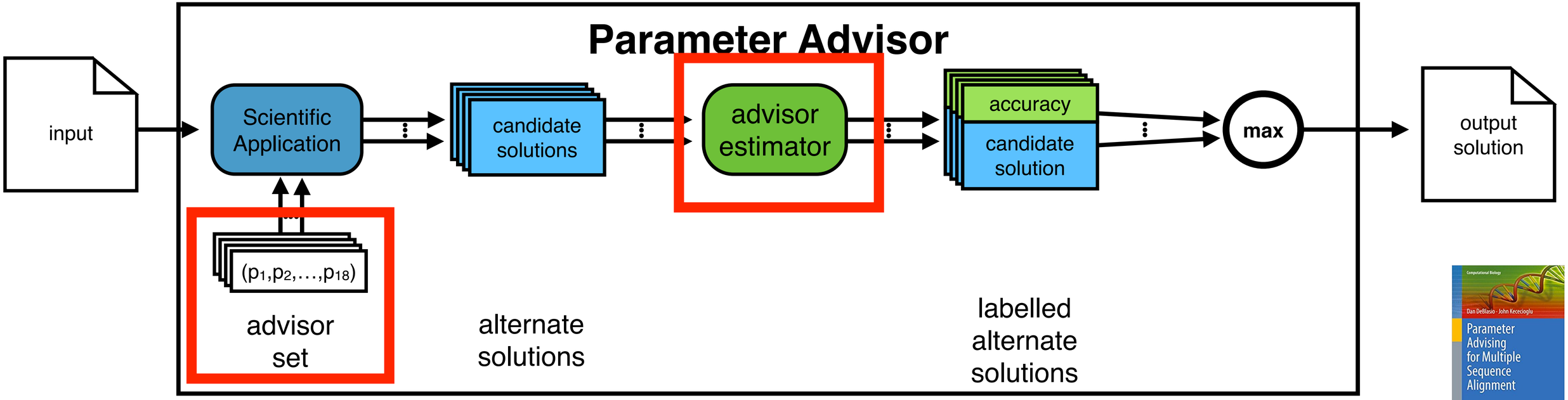
[DK, Springer International, 2017]

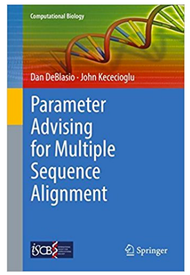
Parameter advising framework

Components of an advisor:

- An advisor set of parameter choice vectors.
- An advisor estimator to rank solutions.

A good advisor set:
- Small
- Representative





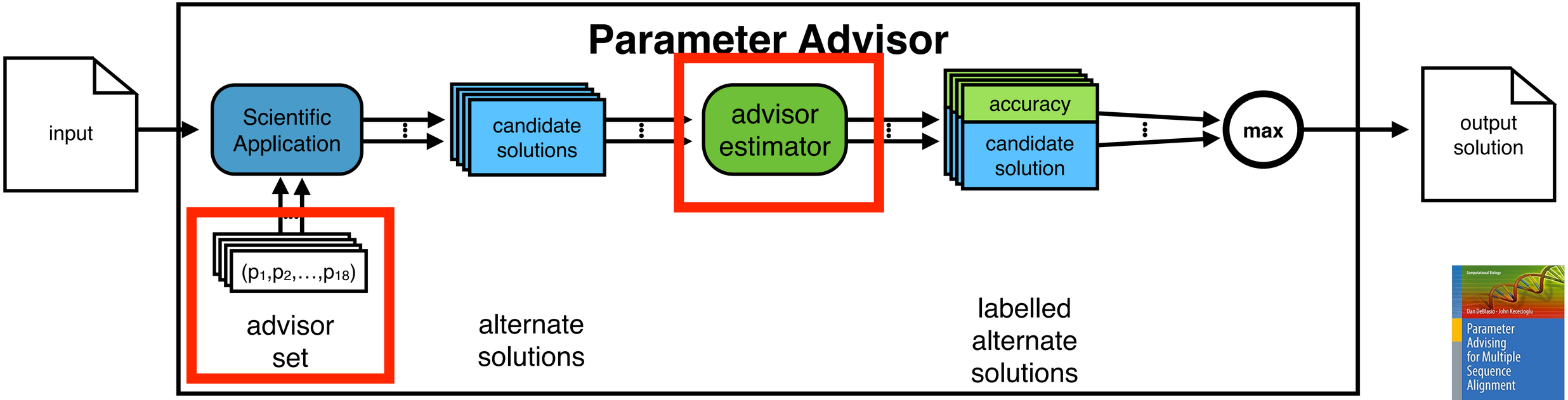
[DK, Springer International, 2017]

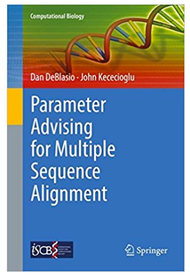
Parameter advising framework

Components of an advisor:

- An advisor set of parameter choice vectors.
- An advisor estimator to rank solutions.

A good advisor estimator:
- Efficient
- Rank Solutions Well



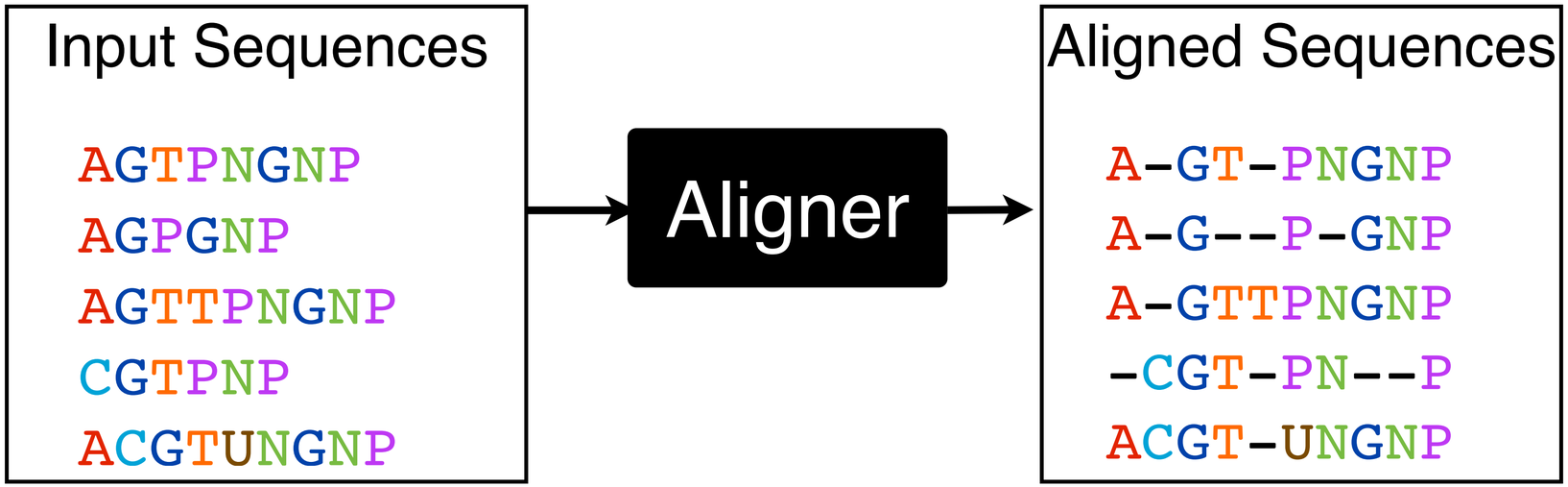


[DK, Springer International, 2017]

Multiple sequence alignment

A fundamental problem in bioinformatics.

- NP-Complete
- many popular aligners
- many parameters whose values affect the output
- no standard metric for measuring accuracy without ground truth



[Wheeler and Kececioglu ISMB 2007]

Accuracy estimation

Alignment accuracy is measured with respect to a reference 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 ···
··· d S R - d ···	··· d S R - - d ···
··· a S H l t ···	··· a S - H l t ···

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 ···
··· 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 ···
··· 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 ···
··· 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 ···
··· 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

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 ···

66% Accuracy

- 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 **E**s**T**imator")

- a polynomial on feature functions
- efficiently learns the coefficients from examples
- uses efficiently computed novel features

[DWK RECOMB2012/DK JCB2012]

Accuracy estimation

Our estimator `Facet` ("**F**eature-based **AC**curacy **E**s**T**imator")

- a polynomial on feature functions
- efficiently learns the coefficients from examples
- uses efficiently computed novel features

Feature functions are the key:
uninformative features → uninformative estimator

[DWK RECOMB2012/DK JCB2012]

Accuracy estimation

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

[DWK RECOMB2012/DK JCB2012]

Accuracy estimation

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

linear estimator

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

[DWK RECOMB2012/DK JCB2012]

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

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

Always linear in the coefficients.

[DWK RECOMB2012/DK JCB2012]

Learning the estimator

We learn the estimator using examples consisting of

- an alignment, and
- its associated true accuracy.

[DWK RECOMB2012/DK JCB2012]

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.

[DWK RECOMB2012/DK JCB2012]

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.

[DWK RECOMB2012/DK JCB2012]

Learning the estimator

$$e_{a,b} \geq E(b) - E(a) = \sum_i c_i \left(f_i(b) - f_i(a)\right)$$

$$e_{a,b} \geq 0$$

$\forall a, b \in$ Examples :

$$Accuracy(a) > Accuracy(b)$$

Learning the estimator

Minimize

$$\sum_{a,b \in \text{examples}} w_{a,b}\, e_{a,b}$$

Subject to:

$$e_{a,b} \geq E(b) - E(a) = \sum_i c_i \left(f_i(b) - f_i(a)\right)$$

$$e_{a,b} \geq 0$$

$\forall a, b \in$ Examples :

$$Accuracy(a) > Accuracy(b)$$

Feature functions

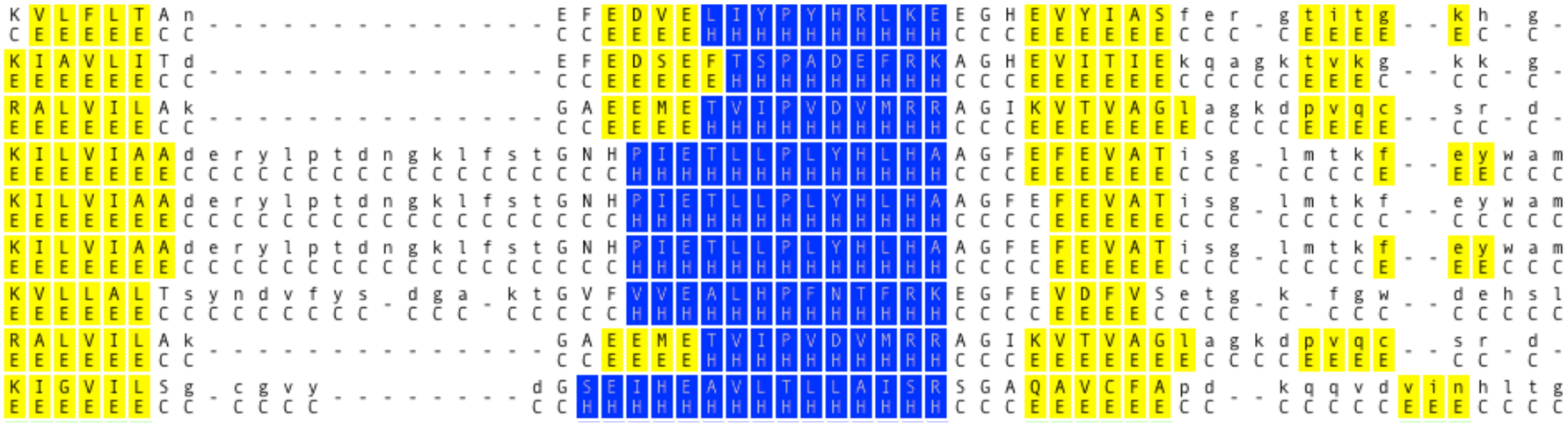
We use protein alignment feature functions that

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

[DWK RECOMB2012/DK JCB2012]

Feature functions

There are three types of secondary structure

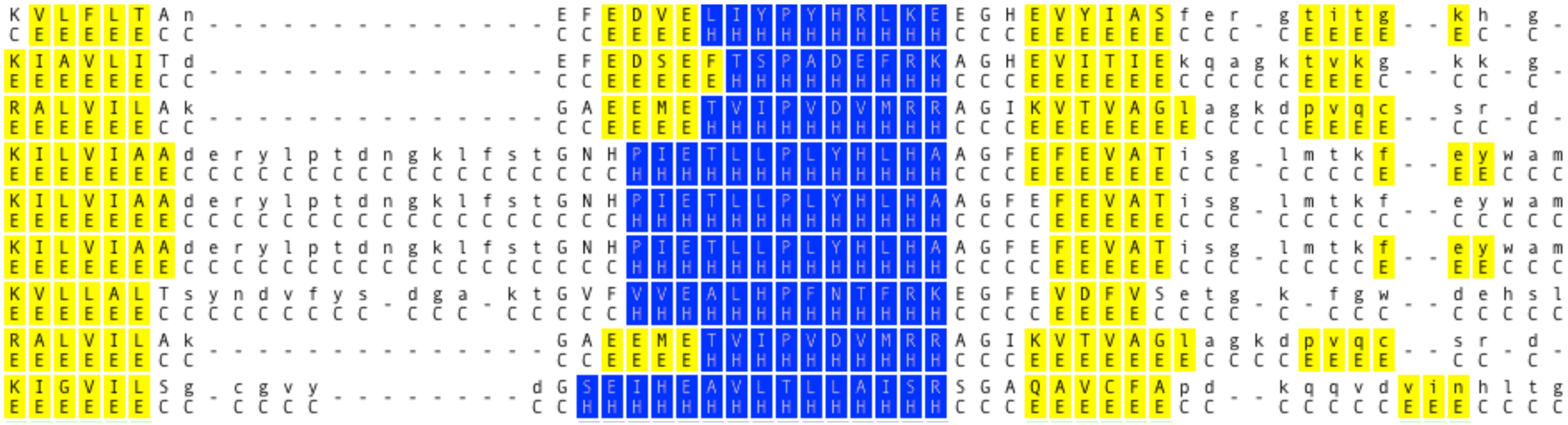


http://www.ebi.ac.uk/training/online/

Feature functions

There are three types of secondary structure

- α-helix,

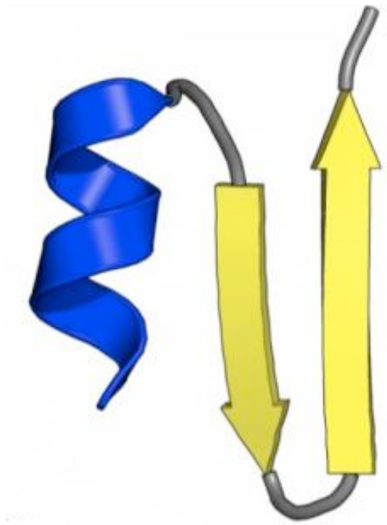


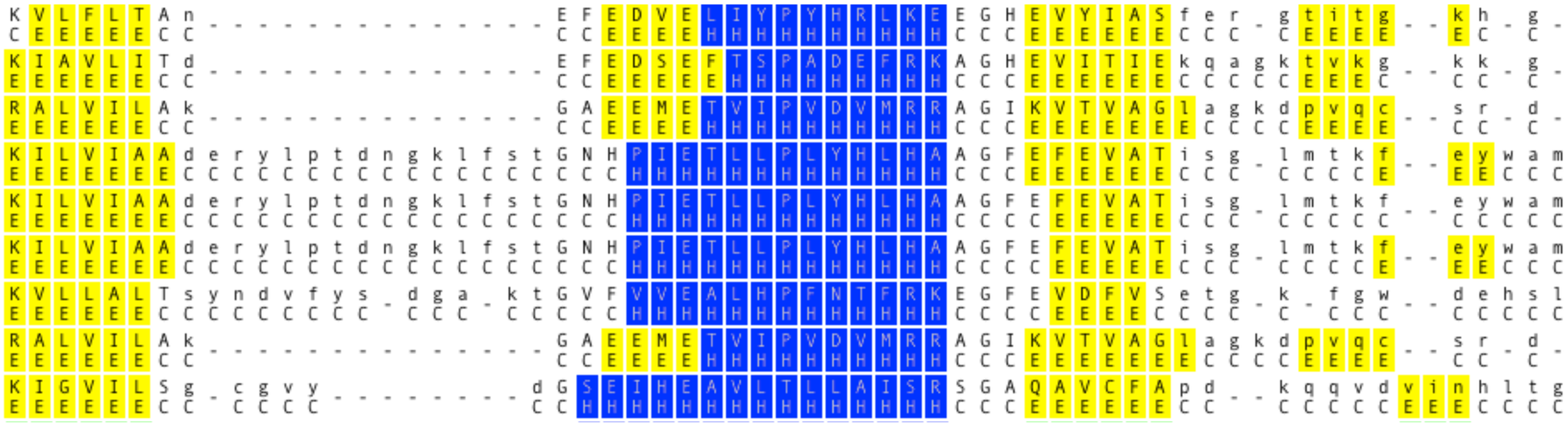
http://www.ebi.ac.uk/training/online/

Feature functions

There are three types of secondary structure

- α-helix,
- β-strand,



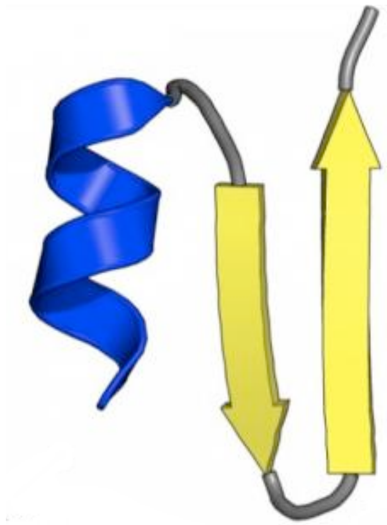


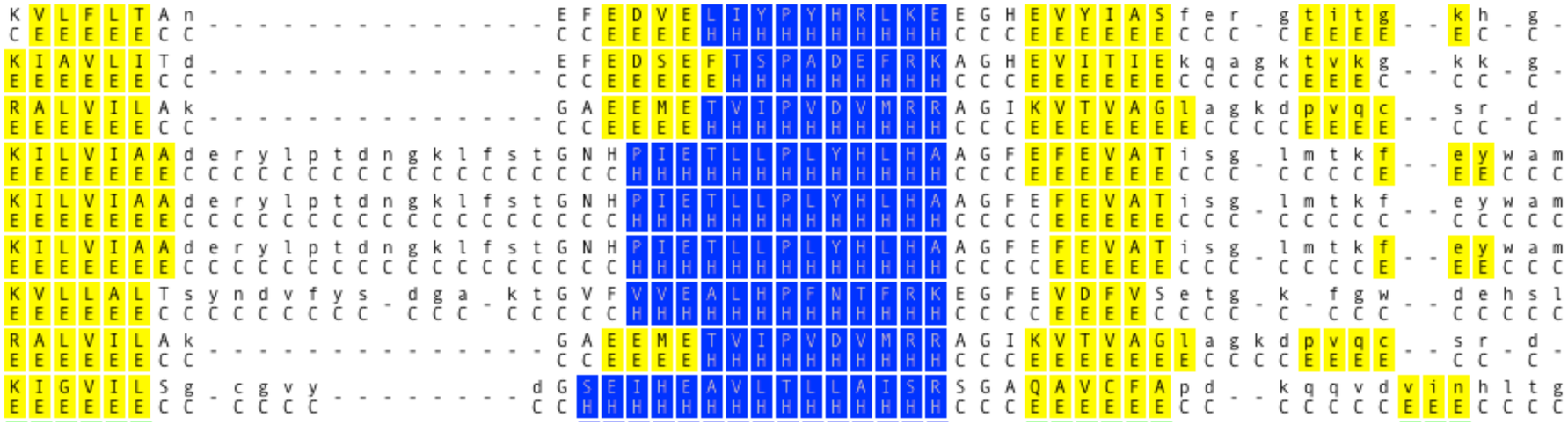
http://www.ebi.ac.uk/training/online/

Feature functions

There are three types of secondary structure

- α-helix,
- β-strand,
- coil.





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

Feature functions

Features based only on the input alignment

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

[DWK RECOMB2012/DK JCB2012]

Feature functions

Features based only on the input alignment

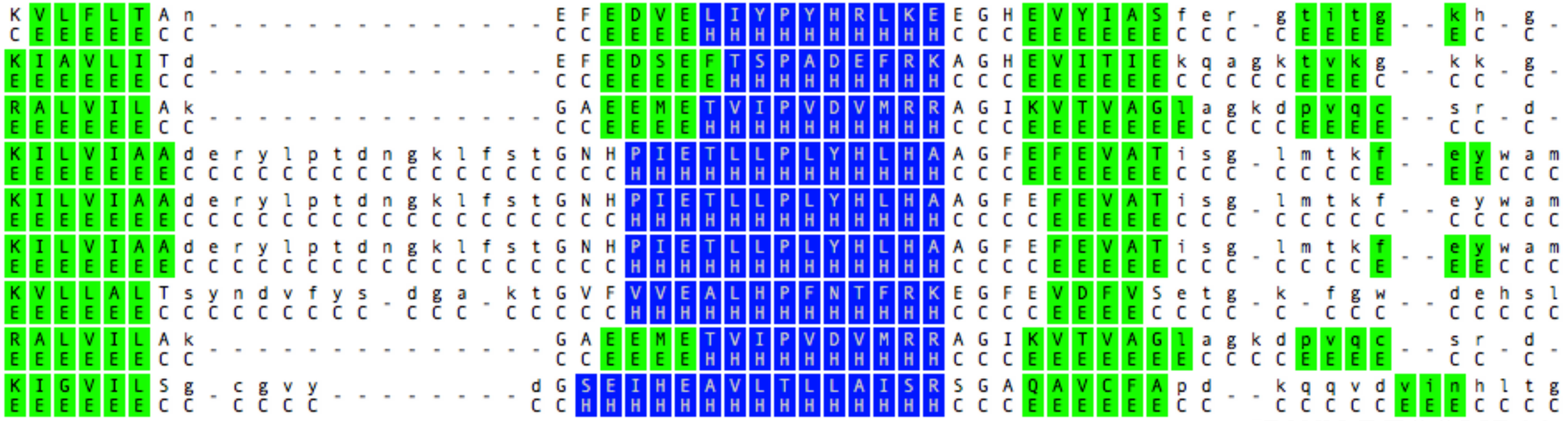
- Amino Acid Identity
- Average Substitution Score
- Information Content
- ...

Features using predicted secondary structure

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

[DWK RECOMB2012/DK JCB2012]

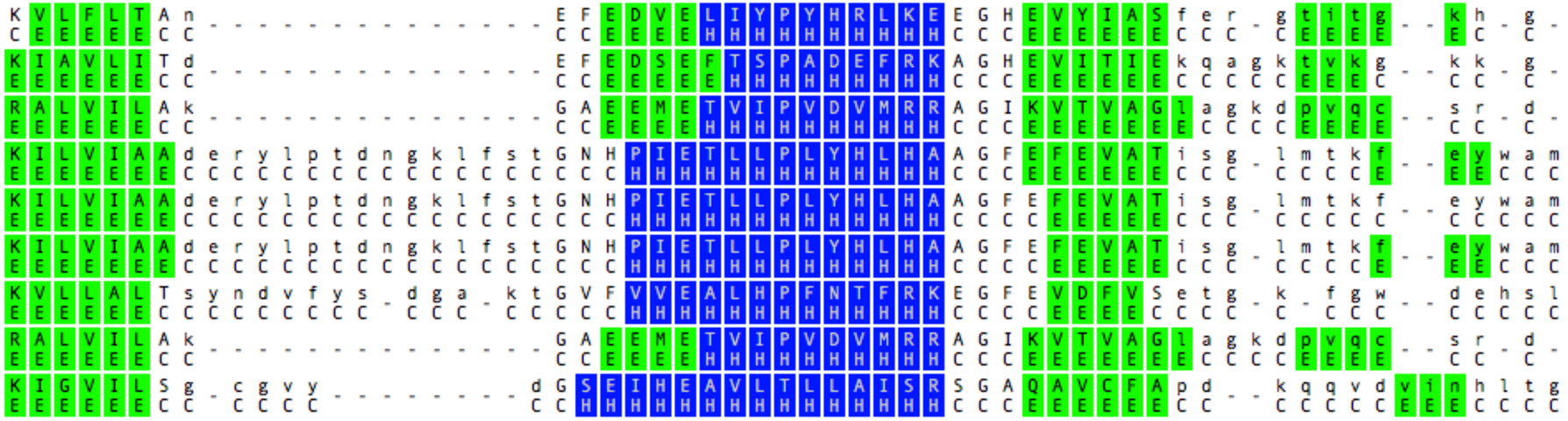
Secondary structure blockiness



[DWK RECOMB2012/DK JCB2012]

Secondary structure blockiness

A block B in alignment A is

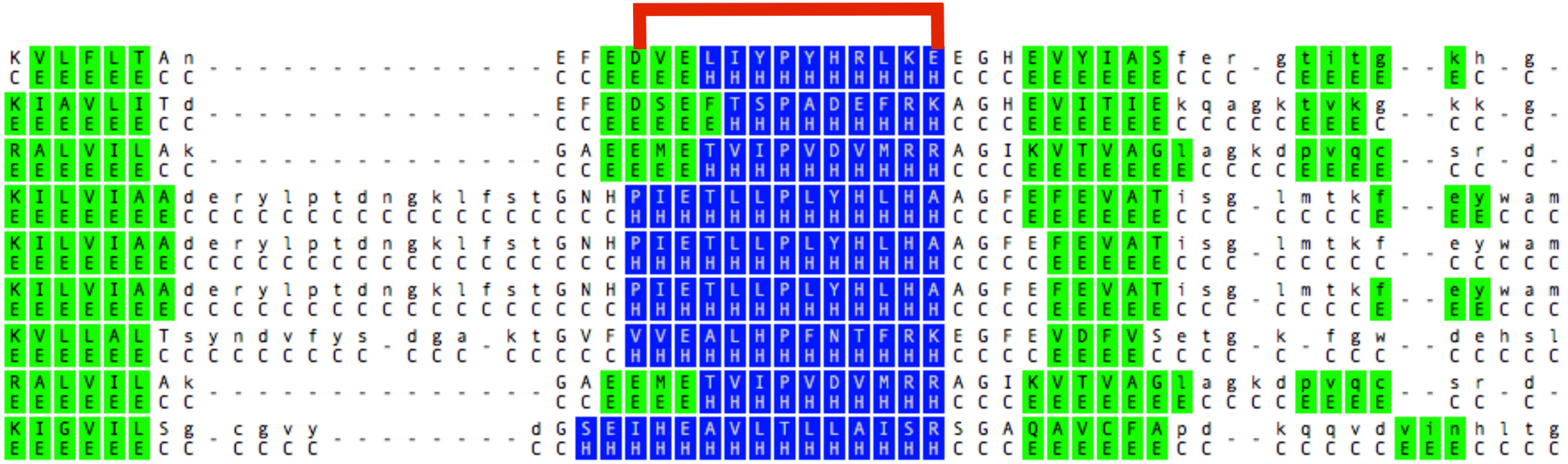


[DWK RECOMB2012/DK JCB2012]

Secondary structure blockiness

A block B in alignment A is

- an interval of at least *l* columns,

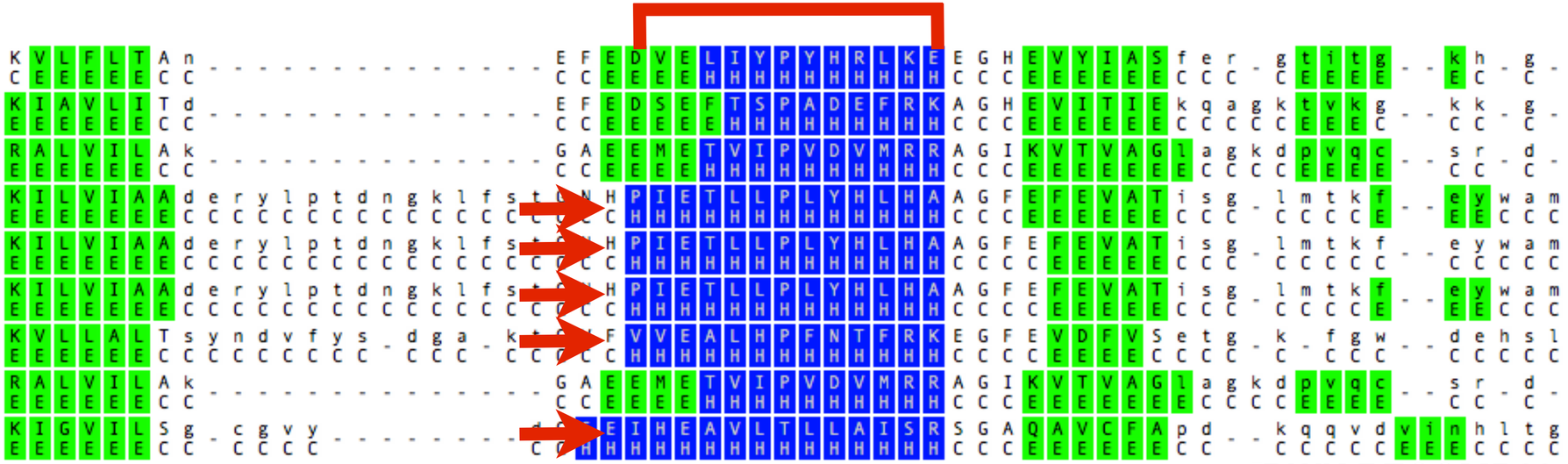


[DWK RECOMB2012/DK JCB2012]

Secondary structure blockiness

A block B in alignment A is

- an interval of at least *l* columns,
- a subset of at least *k* rows,

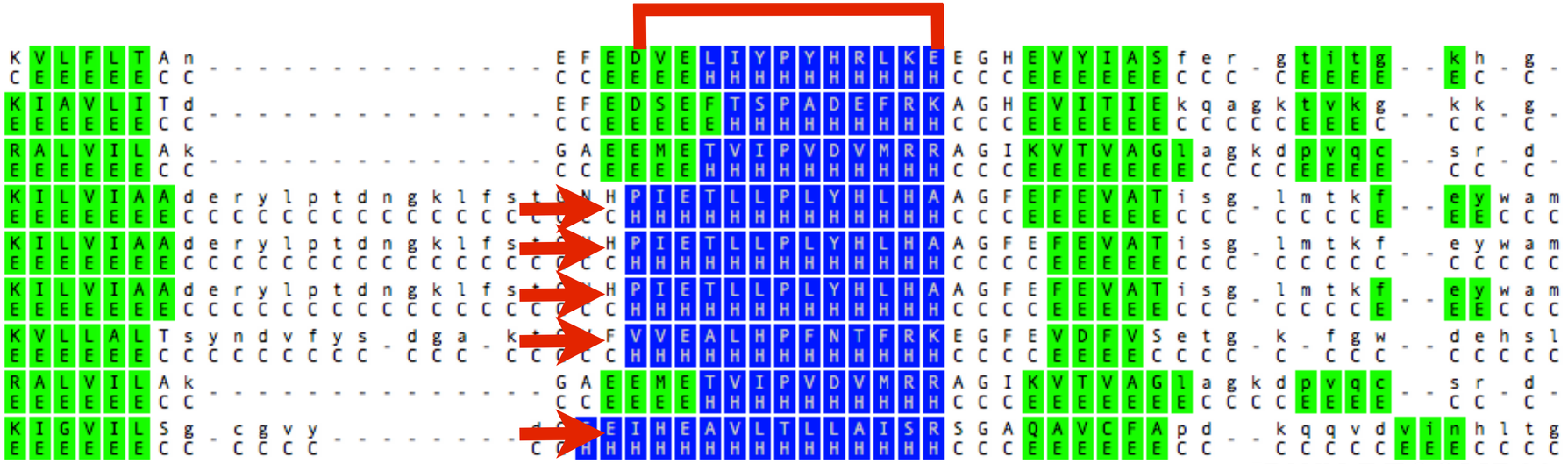


[DWK RECOMB2012/DK JCB2012]

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.

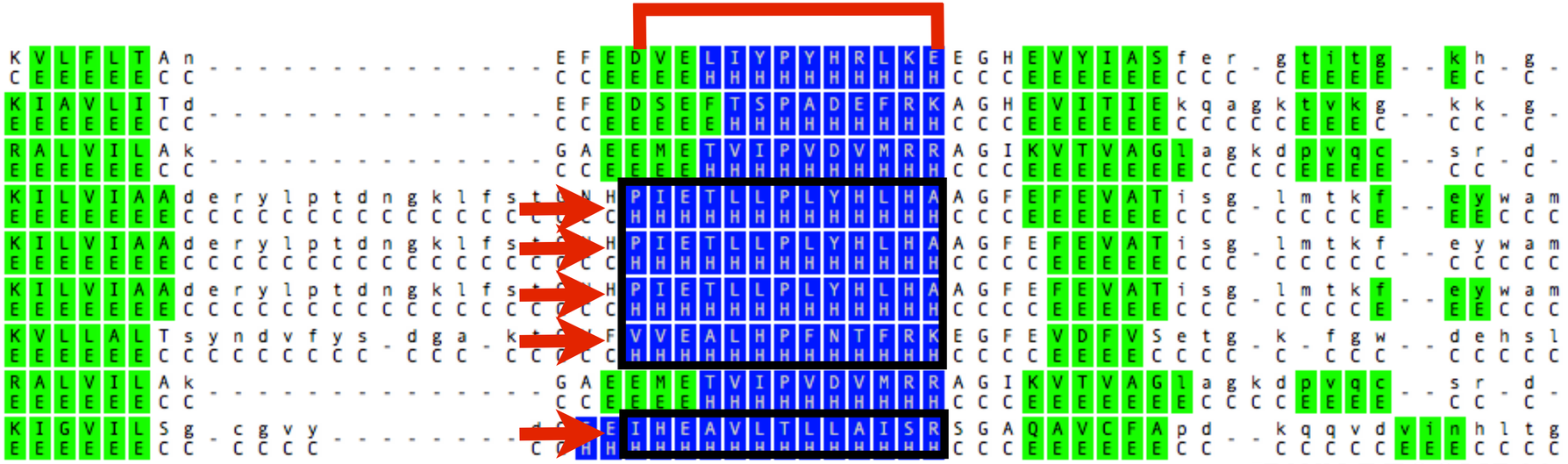


[DWK RECOMB2012/DK JCB2012]

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.

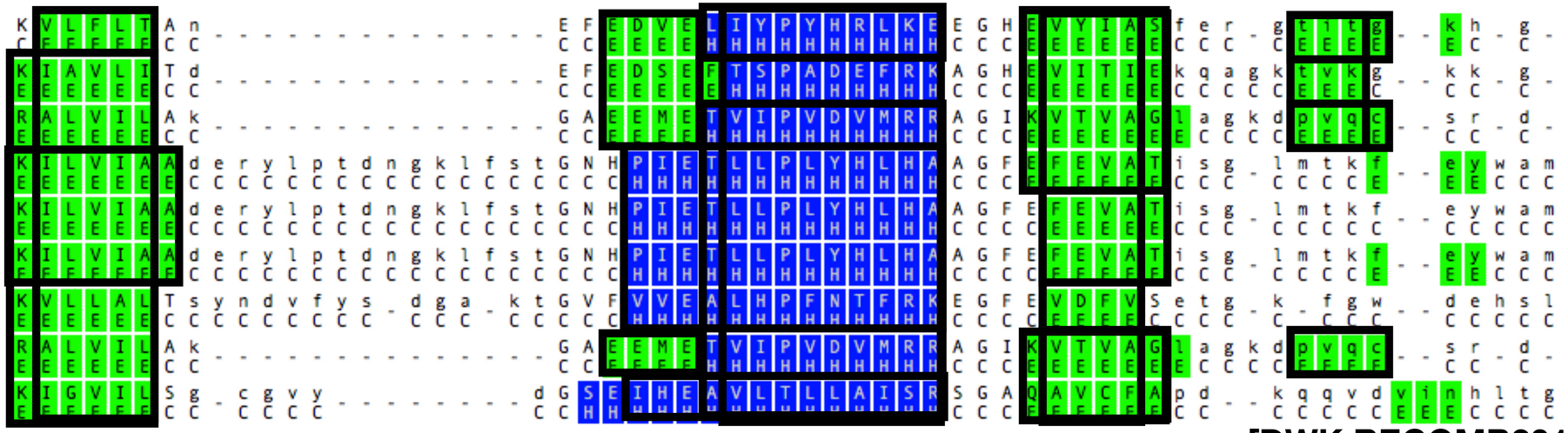


[DWK RECOMB2012/DK JCB2012]

Secondary structure blockiness

A block B in alignment A is

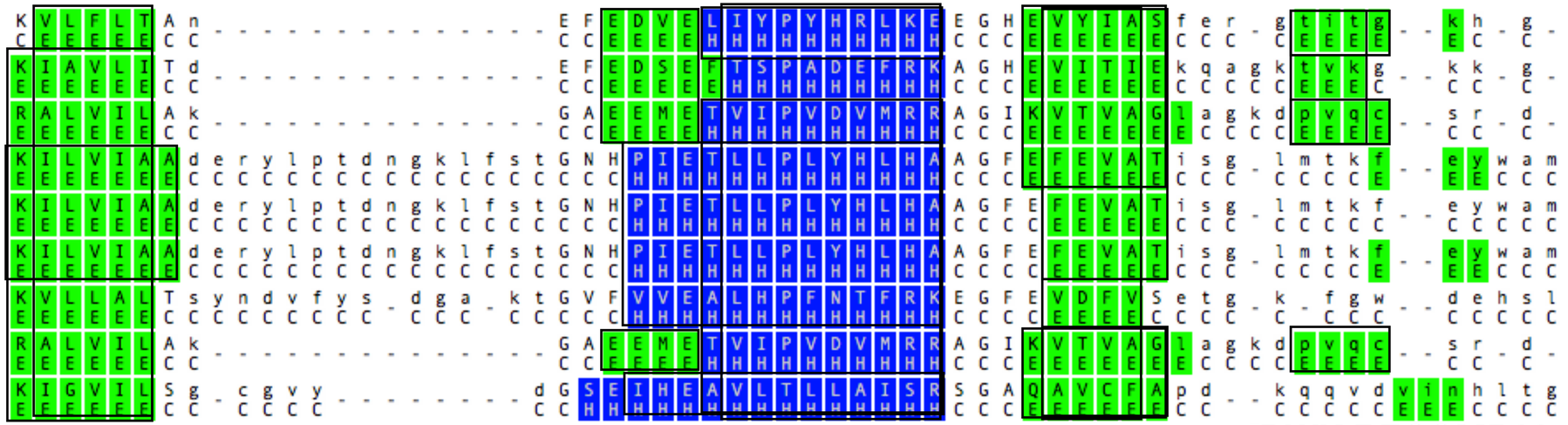
- an interval of at least *l* columns,
- a subset of at least *k* rows,
- with the same secondary structure for all residues in B.



[DWK RECOMB2012/DK JCB2012]

Secondary structure blockiness

A packing P for alignment A is

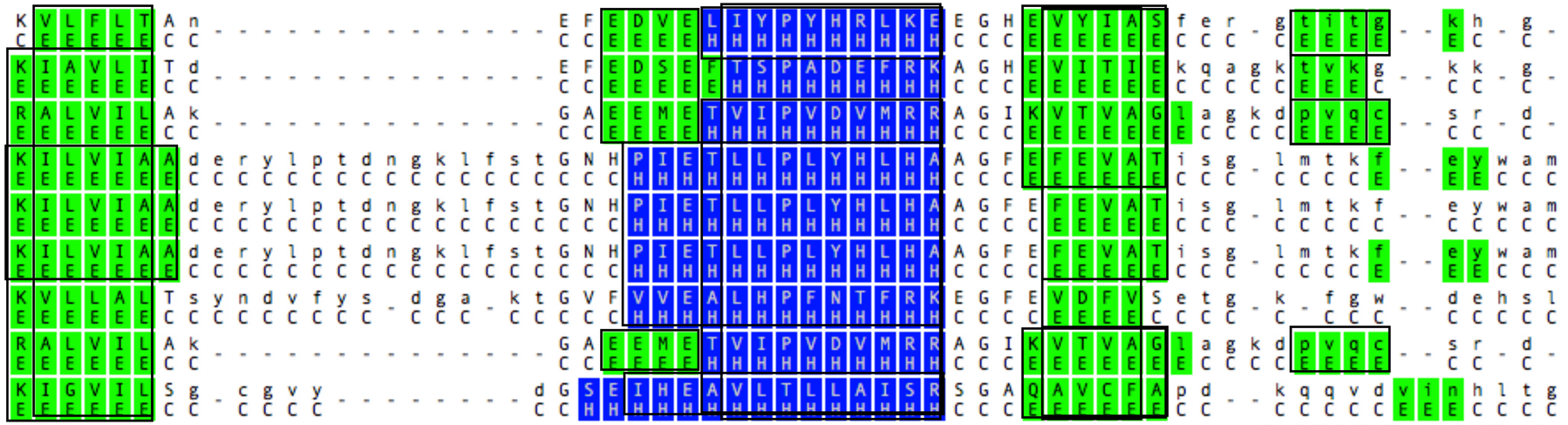


[DWK RECOMB2012/DK JCB2012]

Secondary structure blockiness

A packing P for alignment A is

- a set of blocks from A,

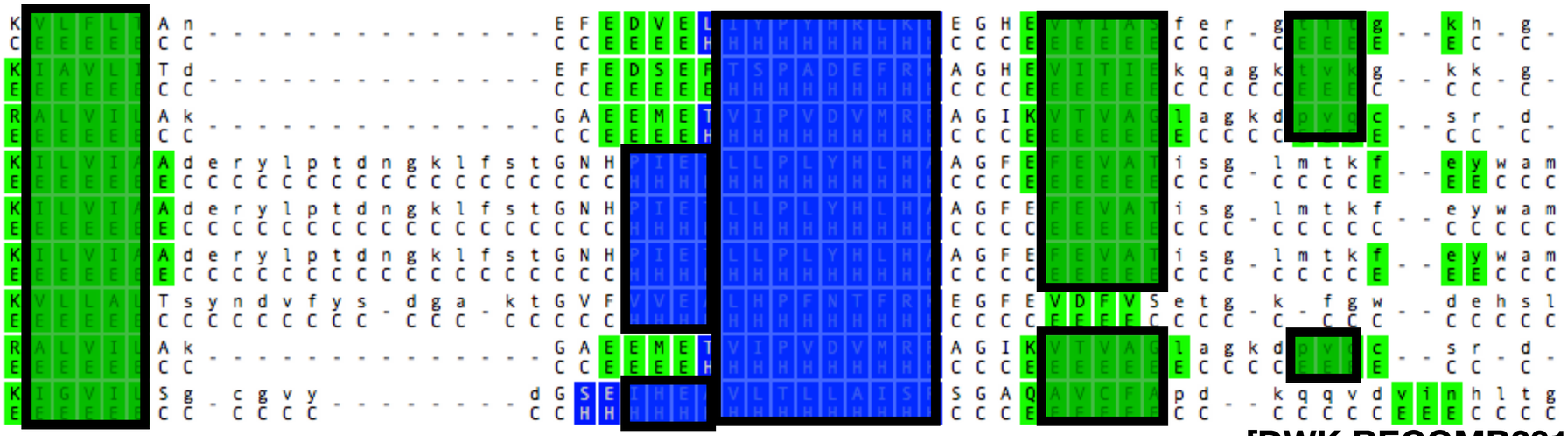


[DWK RECOMB2012/DK JCB2012]

Secondary structure blockiness

A packing P for alignment A is

- a set of blocks from A,
- whose columns are disjoint.



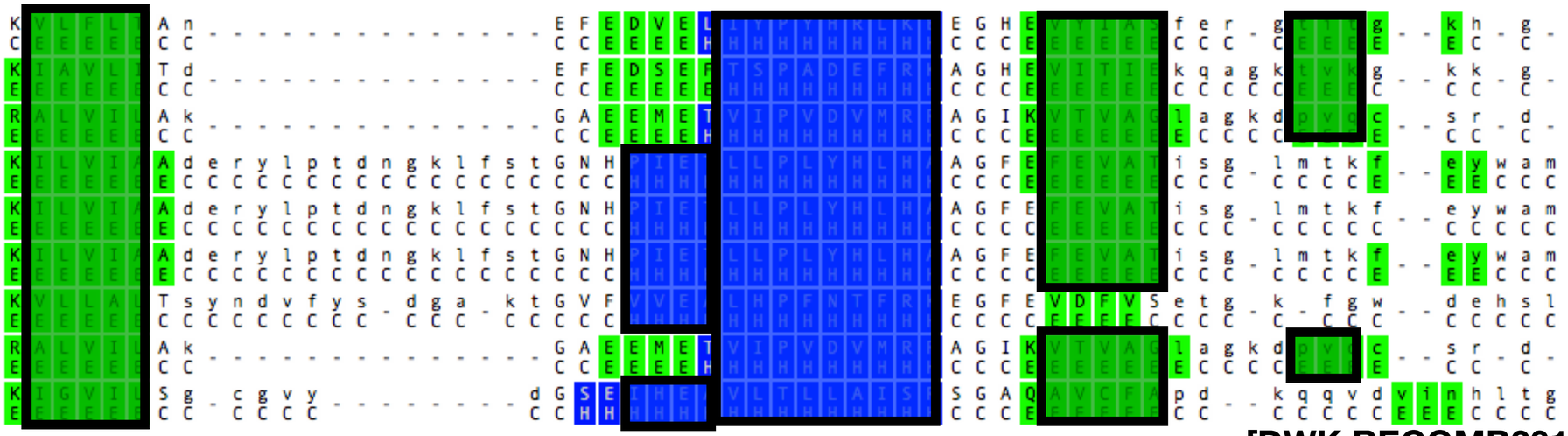
[DWK RECOMB2012/DK JCB2012]

Secondary structure blockiness

A packing P for alignment A is

- a set of blocks from A,
- whose columns are disjoint.

The value of P is the number of substitutions it contains.



[DWK RECOMB2012/DK JCB2012]

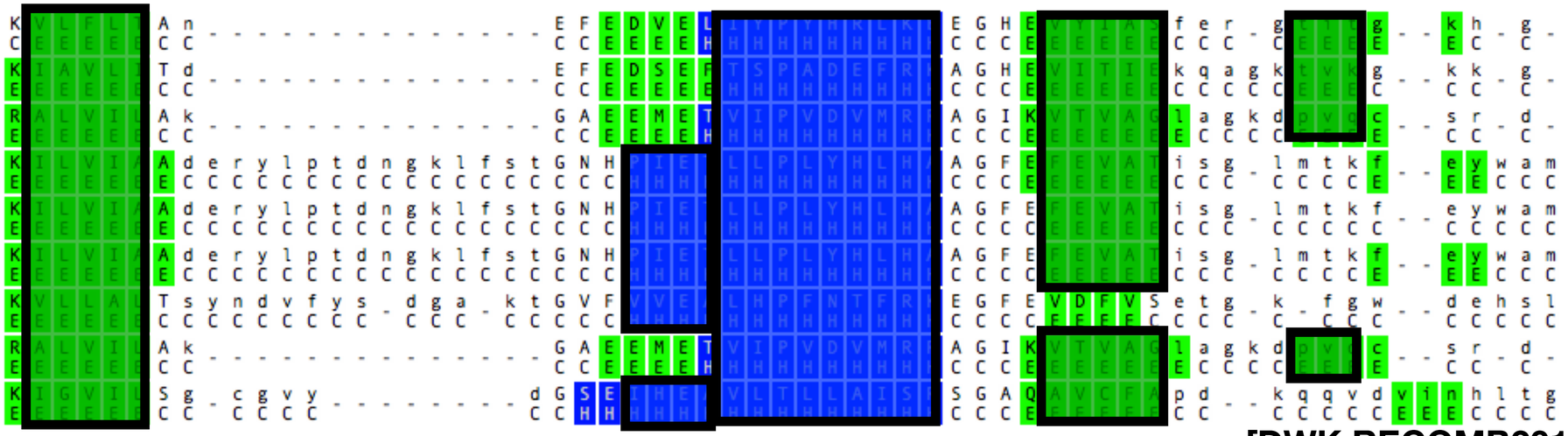
Secondary structure blockiness

A packing P for alignment A is

- a set of blocks from A,
- whose columns are disjoint.

The value of P is the number of substitutions it contains.

The Blockiness feature is the maximum value of any packing.



[DWK RECOMB2012/DK JCB2012]

Secondary structure blockiness

Theorem (Evaluating Blockiness)

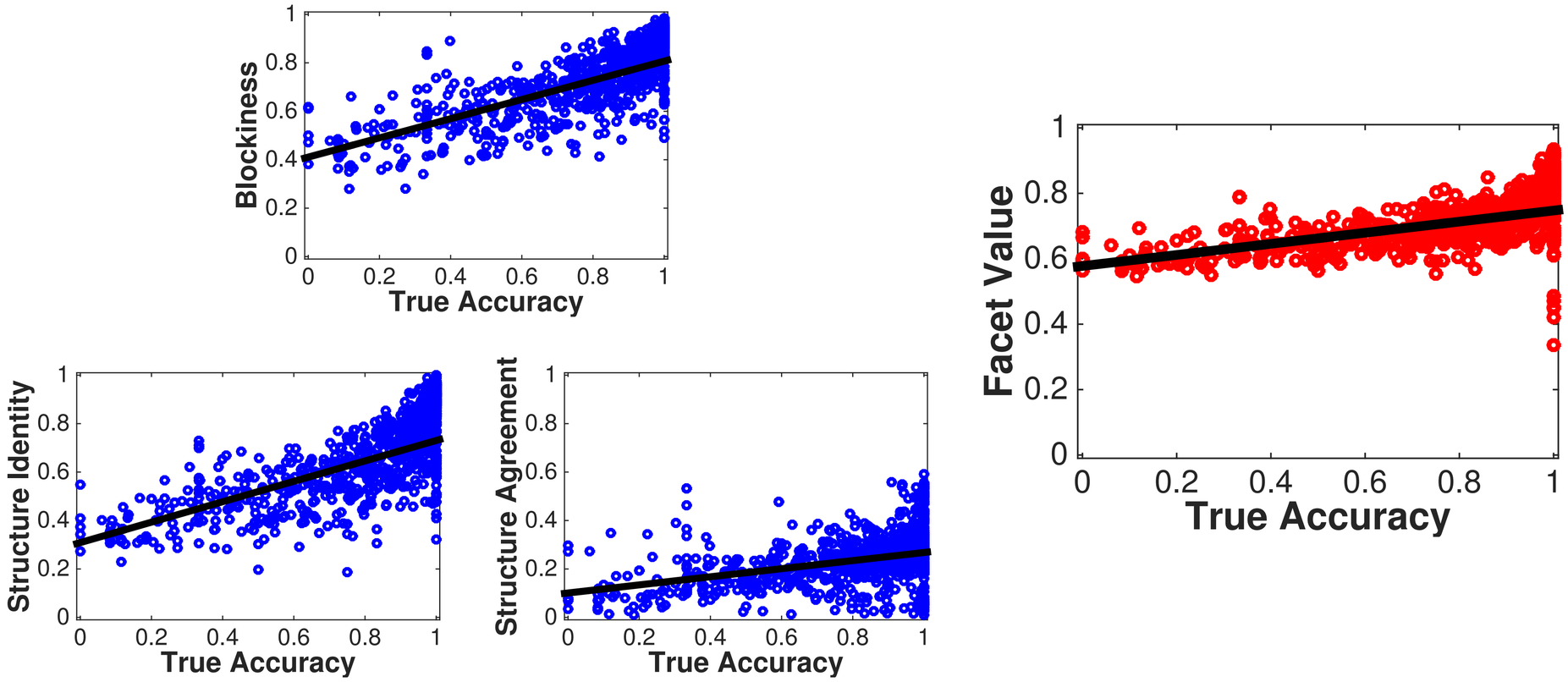
Blockiness can be computed in $O(mn)$ time,
for an alignment with m rows and n columns.

Algorithm translates the problem into finding the longest path in a directed acyclic graph.

[DWK RECOMB2012/DK JCB2012]

Accuracy estimation

Best features trend well with accuracy.

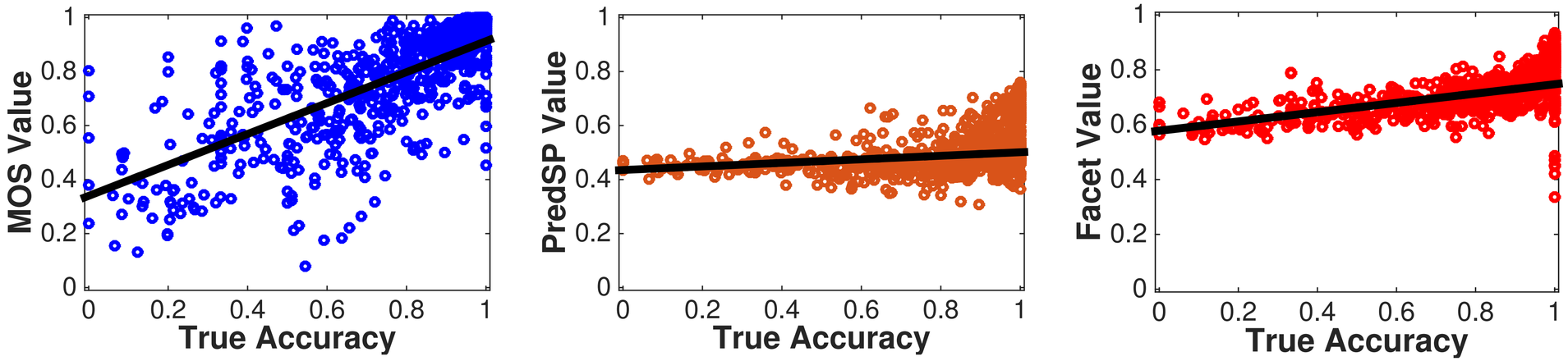


Facet estimator has less spread than its features.

[DWK RECOMB2012/DK JCB2012]

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

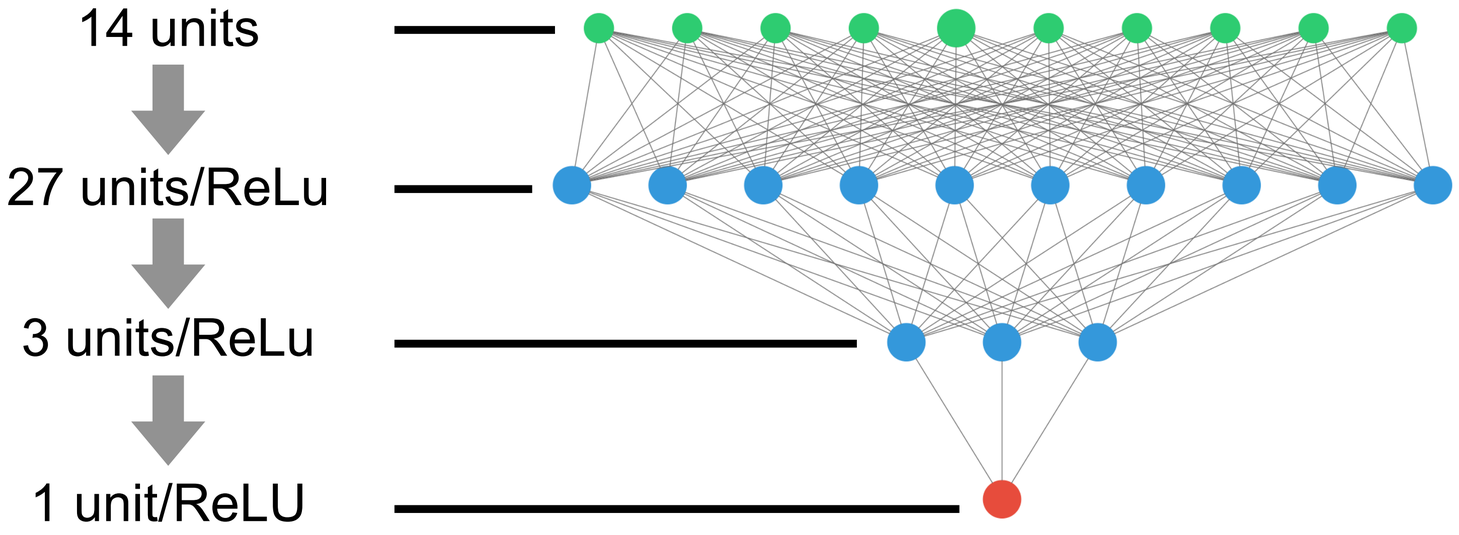
Facet's slope and spread is best for advising

[MOS: Lassmann and Sonnhammer, NAR 2005] [PredSP: Ahola, et al. Bioinformatics 2008] [DWK RECOMB2012/DK JCB2012]

Exploiting non-linearity

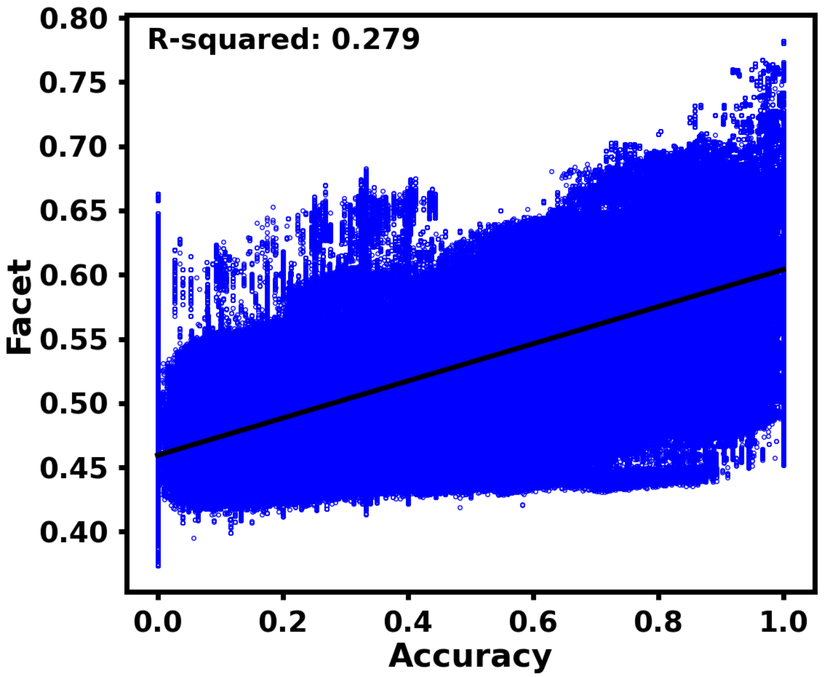
While we designed the features to scale linearly with accuracy, some show some non-linear behavior when plotted.

- Advanced machine learning allowed for the use of a neural network predictor.
- We also produced a much larger training set (now >14M alignments).

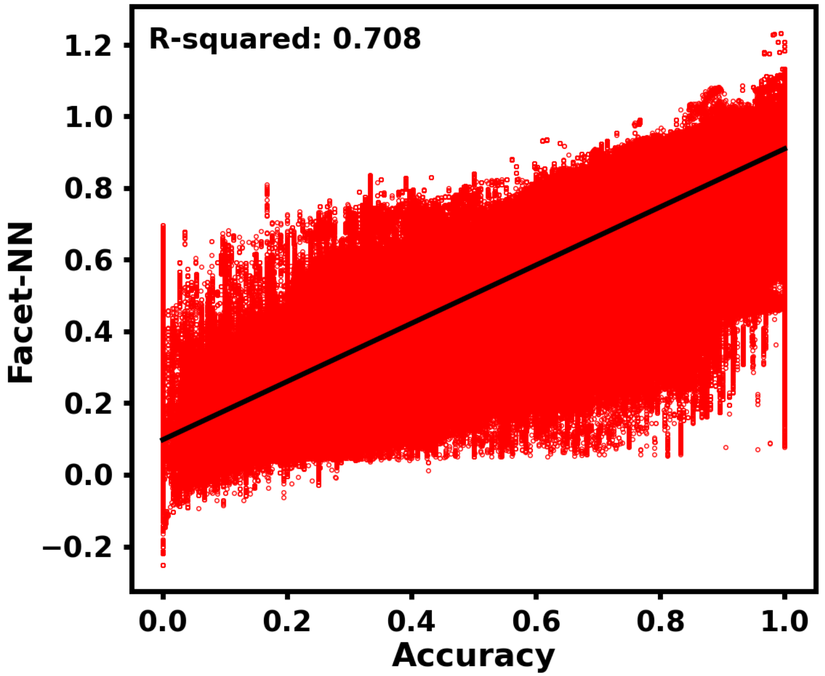


Exploiting non-linearity

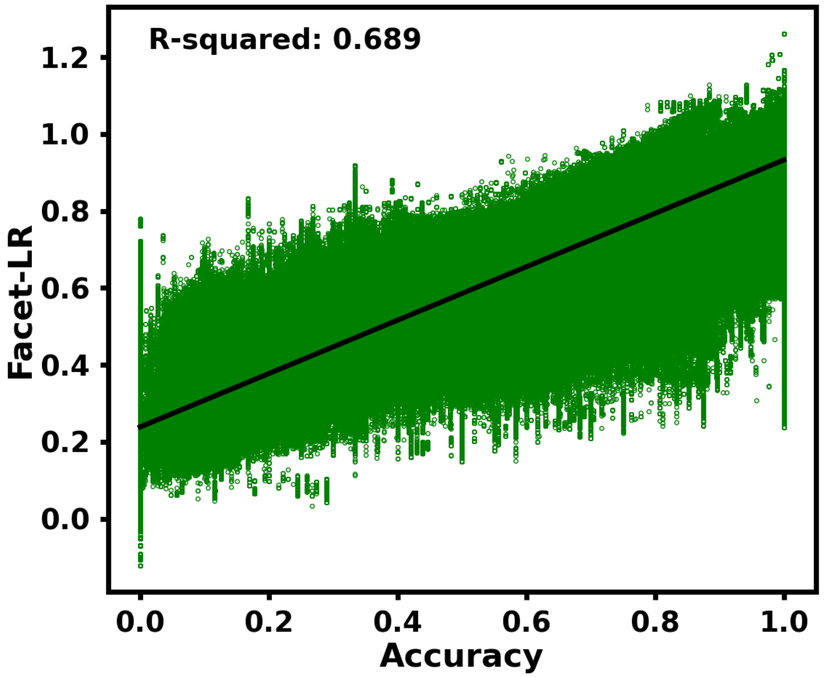
Previous Result



Neural Network

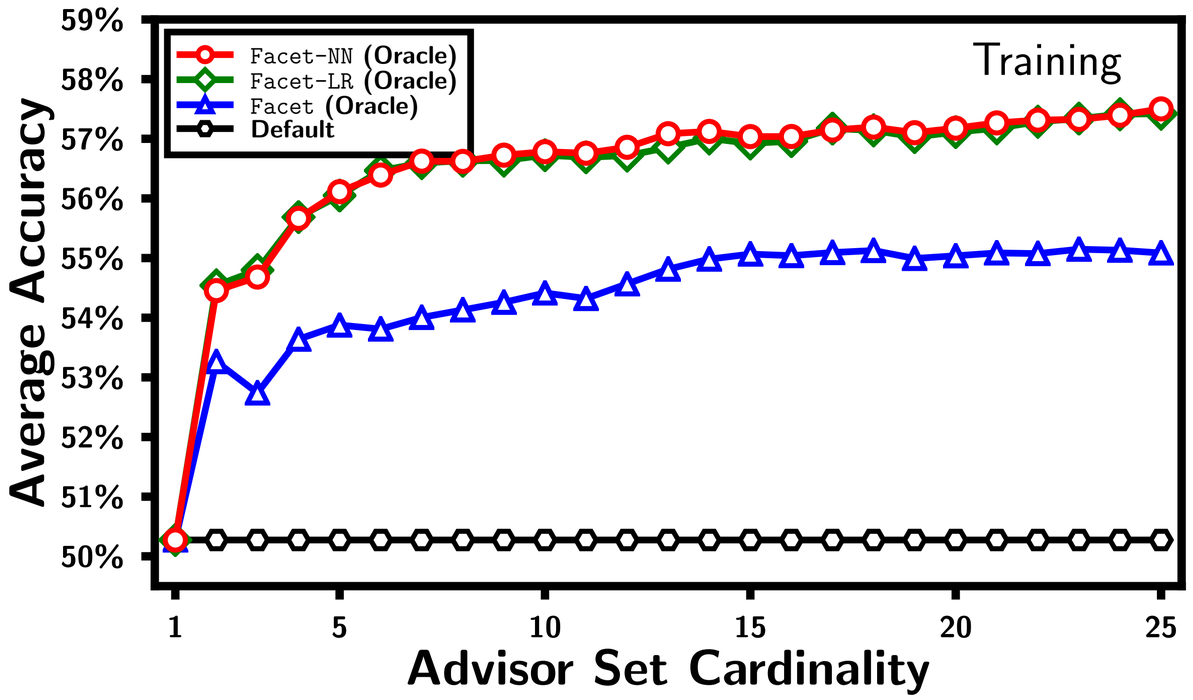


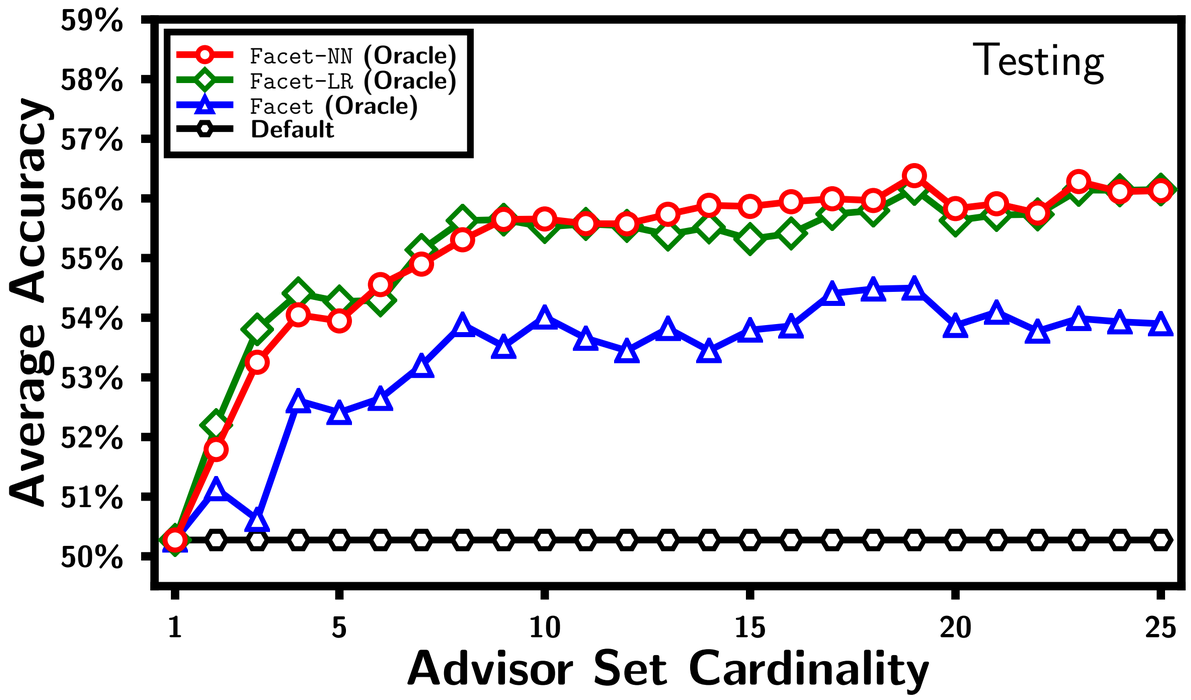
Linear Regression



Modern techniques and larger training also lead to a more accurate linear model.

Advising for Multiple Sequence Alignment



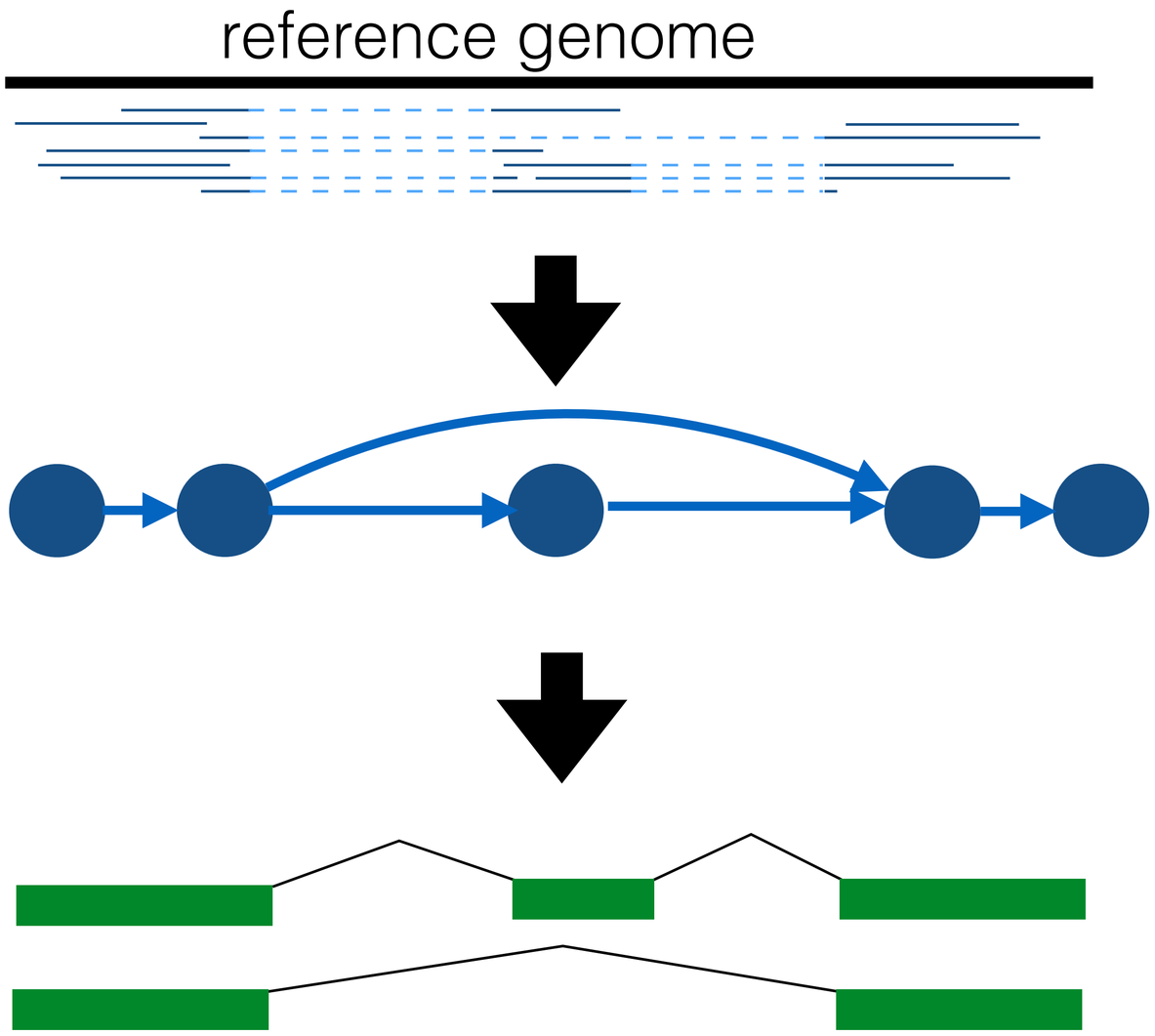


Facet-NN and Facet-LR outperform original Facet on the advising task.

Transcript assembly

TA is fundamental in transcriptomics.

- It's computationally difficult.
- It's easily impacted by choices of parameter values.
- There is no readily available way to confirm an assembly's accuracy.



[Shao and Kingsford, Nature Biotechnology 2017]

Transcript assembly

For the human genome there is a reference transcriptome.

- Contains a large set of biologically verified transcripts.
- More than will be seen in a single experiment.
- Missing novel transcripts for any given experiment.

Transcript assembly

For the human genome there is a reference transcriptome.

- Contains a large set of biologically verified transcripts.
- More than will be seen in a single experiment.
- Missing novel transcripts for any given experiment.

Area Under the Curve (AUC) can be calculated using the reference transcriptome.

- Map assembled transcripts to the reference.

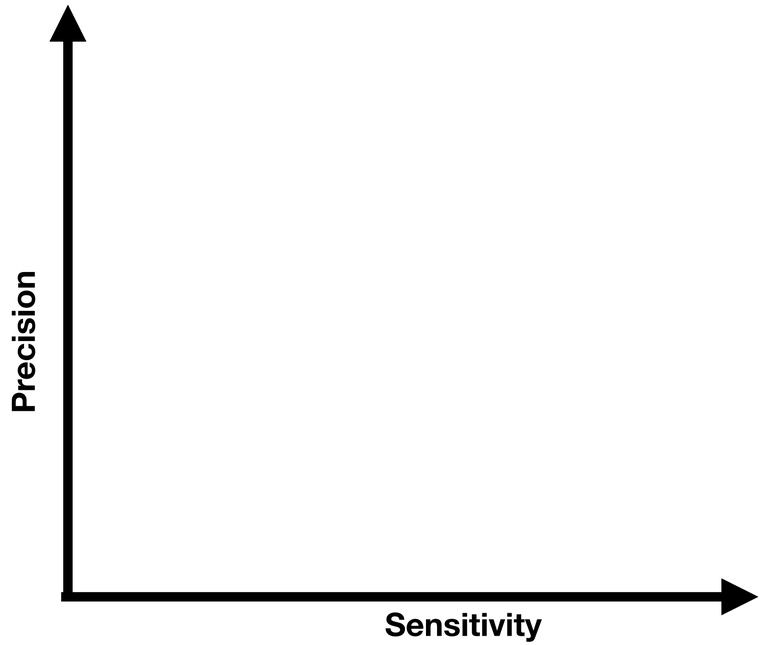
Transcript assembly

For the human genome there is a reference transcriptome.

- Contains a large set of biologically verified transcripts.
- More than will be seen in a single experiment.
- Missing novel transcripts for any given experiment.

Area Under the Curve (AUC) can be calculated using the reference transcriptome.

- Map assembled transcripts to the reference.
- Threshold the quality score from the assembler to get precision/sensitivity.



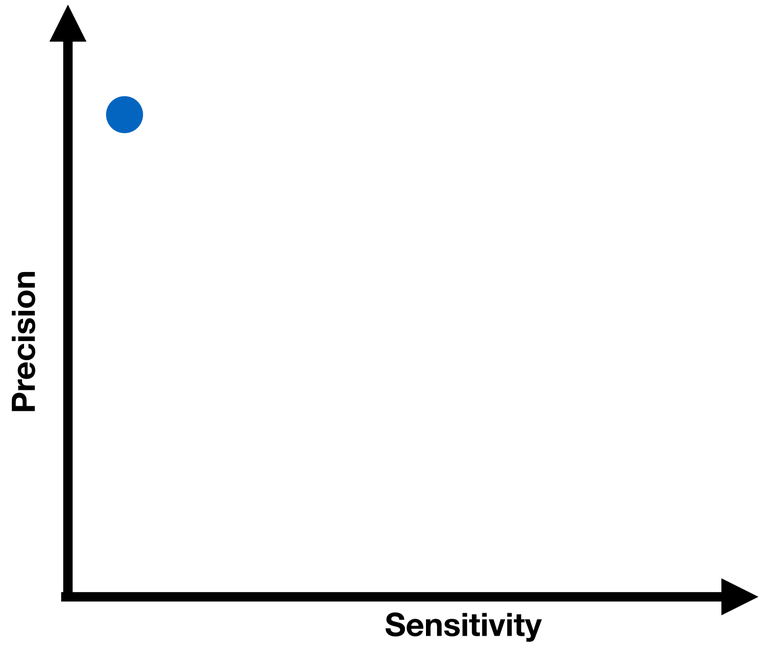
Transcript assembly

For the human genome there is a reference transcriptome.

- Contains a large set of biologically verified transcripts.
- More than will be seen in a single experiment.
- Missing novel transcripts for any given experiment.

Area Under the Curve (AUC) can be calculated using the reference transcriptome.

- Map assembled transcripts to the reference.
- Threshold the quality score from the assembler to get precision/sensitivity.



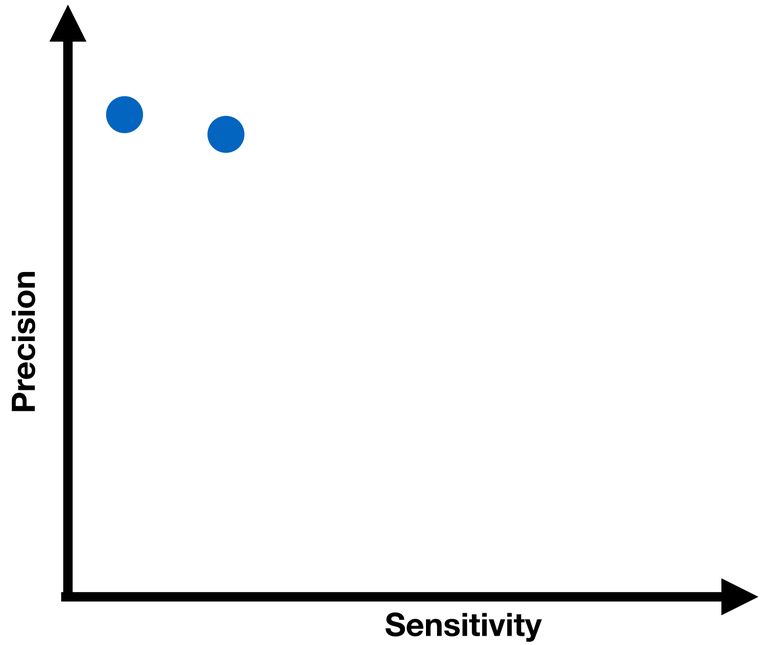
Transcript assembly

For the human genome there is a reference transcriptome.

- Contains a large set of biologically verified transcripts.
- More than will be seen in a single experiment.
- Missing novel transcripts for any given experiment.

Area Under the Curve (AUC) can be calculated using the reference transcriptome.

- Map assembled transcripts to the reference.
- Threshold the quality score from the assembler to get precision/sensitivity.



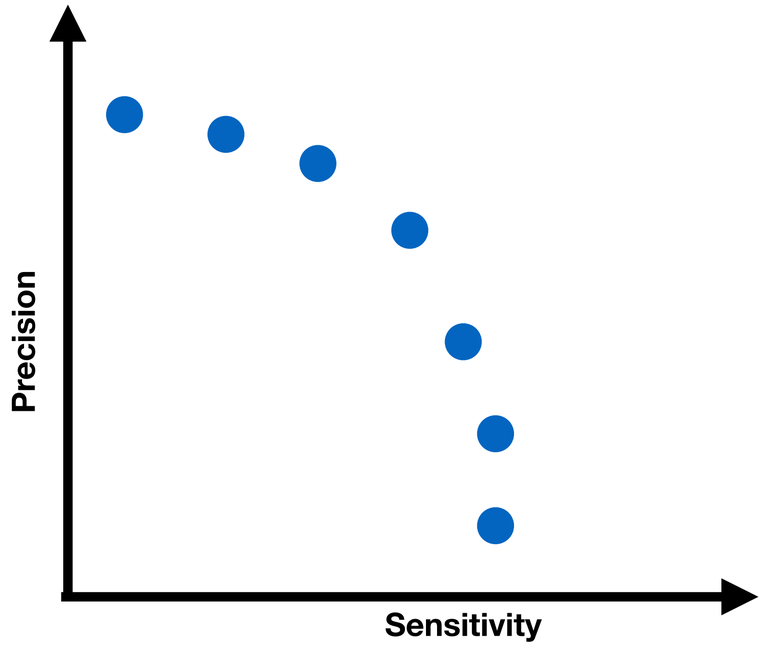
Transcript assembly

For the human genome there is a reference transcriptome.

- Contains a large set of biologically verified transcripts.
- More than will be seen in a single experiment.
- Missing novel transcripts for any given experiment.

Area Under the Curve (AUC) can be calculated using the reference transcriptome.

- Map assembled transcripts to the reference.
- Threshold the quality score from the assembler to get precision/sensitivity.



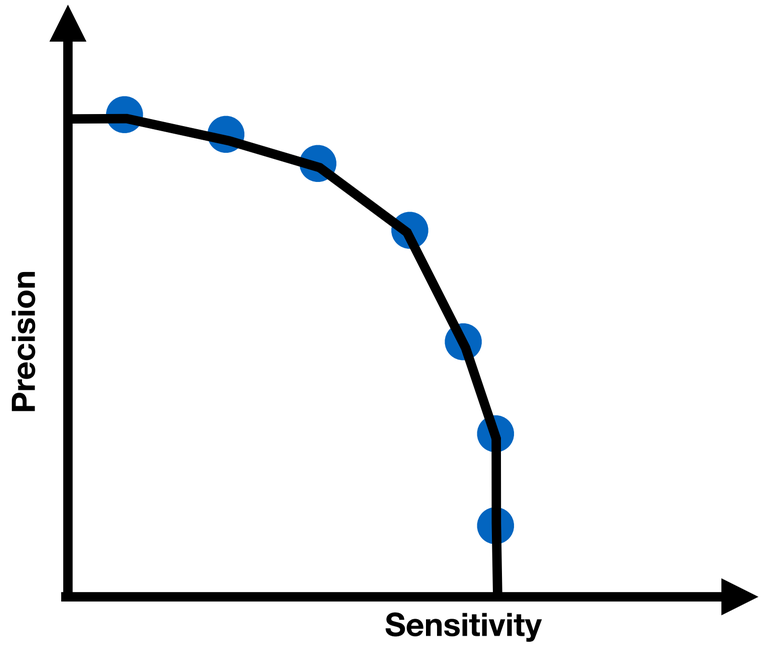
Transcript assembly

For the human genome there is a reference transcriptome.

- Contains a large set of biologically verified transcripts.
- More than will be seen in a single experiment.
- Missing novel transcripts for any given experiment.

Area Under the Curve (AUC) can be calculated using the reference transcriptome.

- Map assembled transcripts to the reference.
- Threshold the quality score from the assembler to get precision/sensitivity.



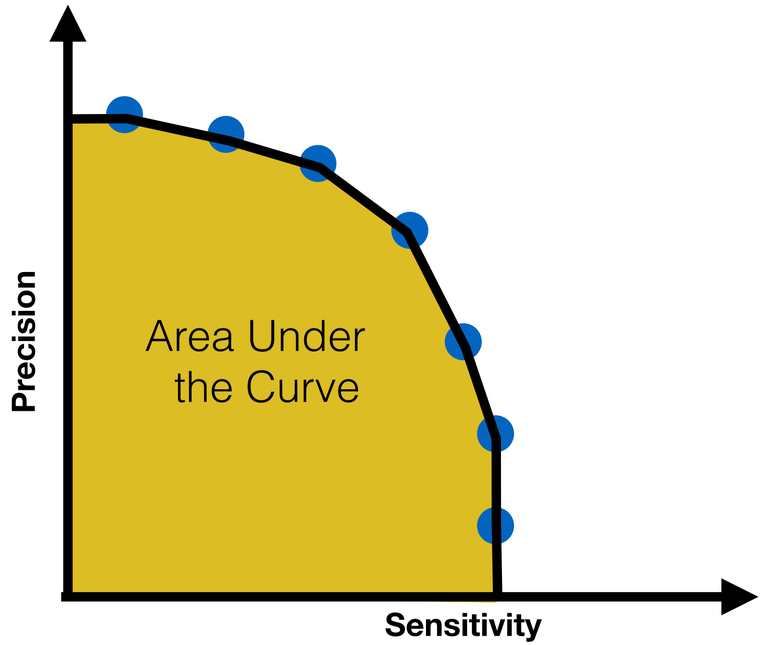
Transcript assembly

For the human genome there is a reference transcriptome.

- Contains a large set of biologically verified transcripts.
- More than will be seen in a single experiment.
- Missing novel transcripts for any given experiment.

Area Under the Curve (AUC) can be calculated using the reference transcriptome.

- Map assembled transcripts to the reference.
- Threshold the quality score from the assembler to get precision/sensitivity.



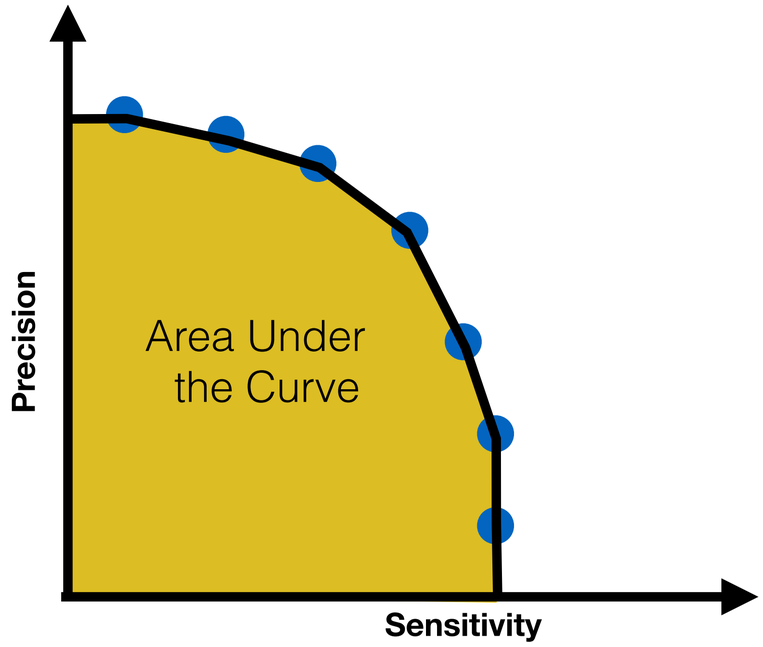
Transcript assembly

For the human genome there is a reference transcriptome.

- Contains a large set of biologically verified transcripts.
- More than will be seen in a single experiment.
- Missing novel transcripts for any given experiment.

Area Under the Curve (AUC) can be calculated using the reference transcriptome.

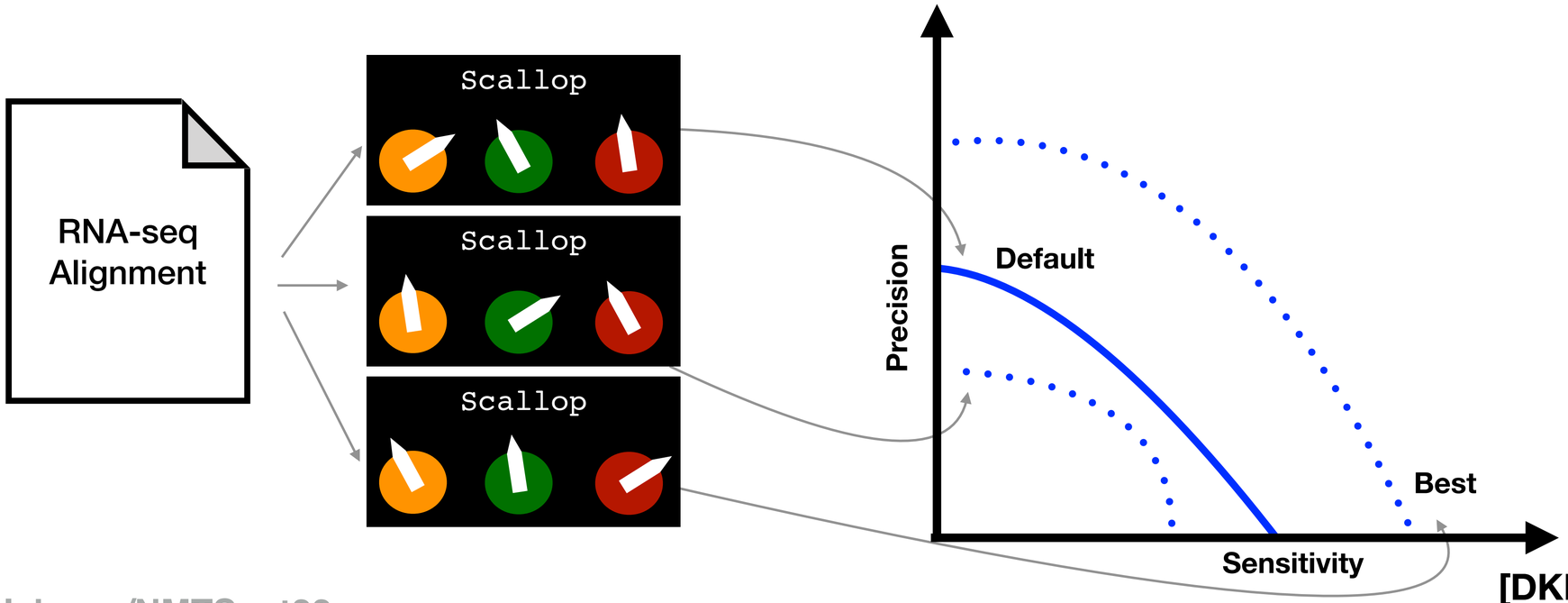
- Map assembled transcripts to the reference.
- Threshold the quality score from the assembler to get precision/sensitivity.
- Commonly used to compare assembler quality.



Transcript assembly advising

Advisor estimator:

- area under the curve



[DKK, WCB@ICML 2019]

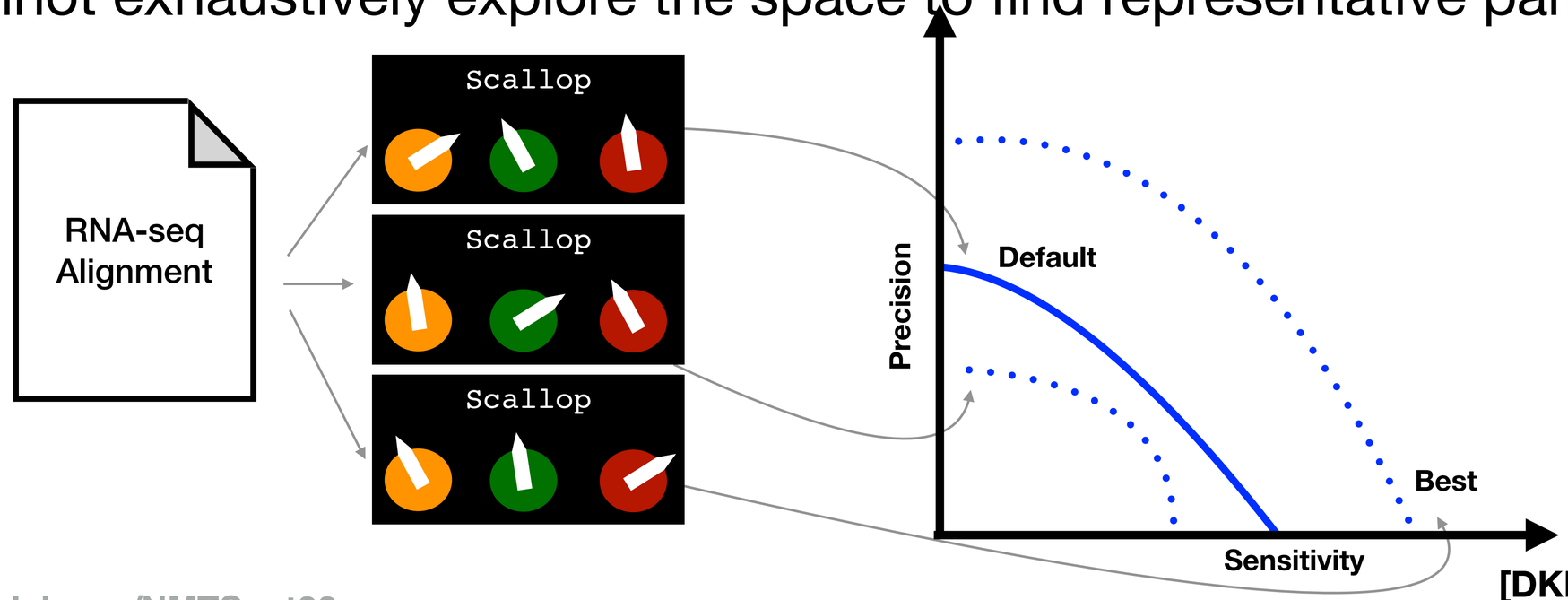
Transcript assembly advising

Advisor estimator:

- area under the curve

Advisor set:

- the number of tunable parameters is very large
- cannot exhaustively explore the space to find representative parameter vectors

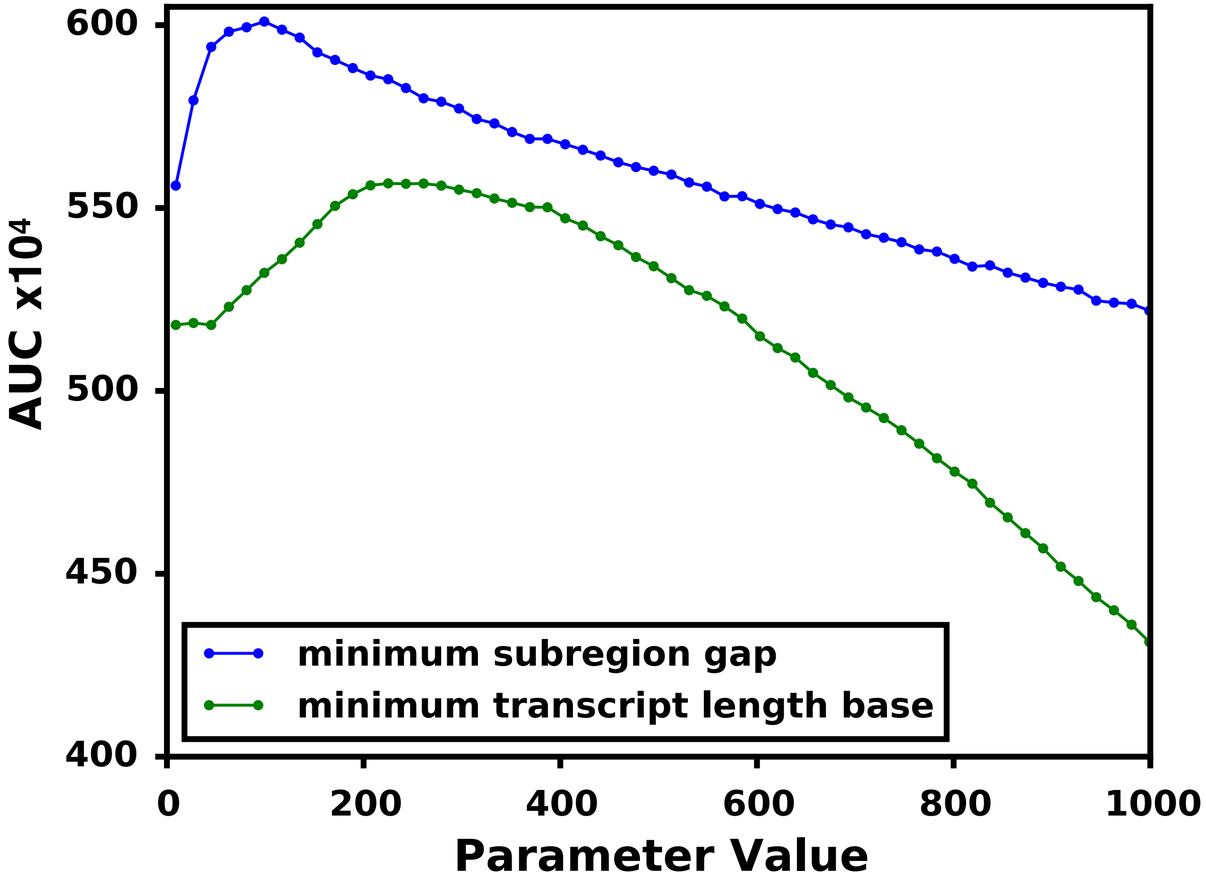


[DKK, WCB@ICML 2019]

Finding an advisor set

Use information about parameter behavior to guide advisor set construction.

- Tested the influence of each parameter.
- Single maximum in the regions tested.

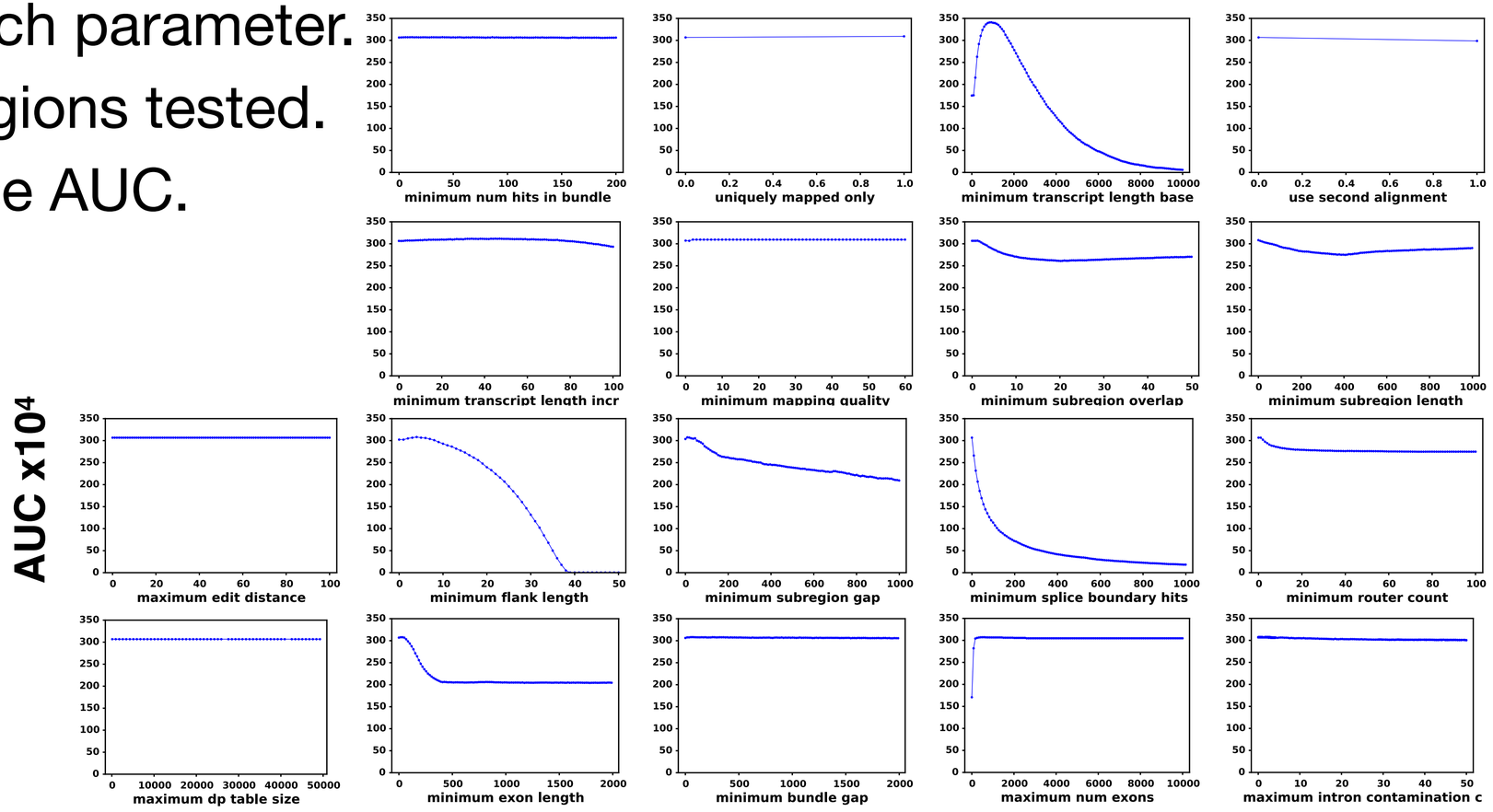


[DKK, WCB@ICML 2019]

Finding an advisor set

Use information about parameter behavior to guide advisor set construction.

- Tested the influence of each parameter.
- Single maximum in the regions tested.
- Many parameters influence AUC.

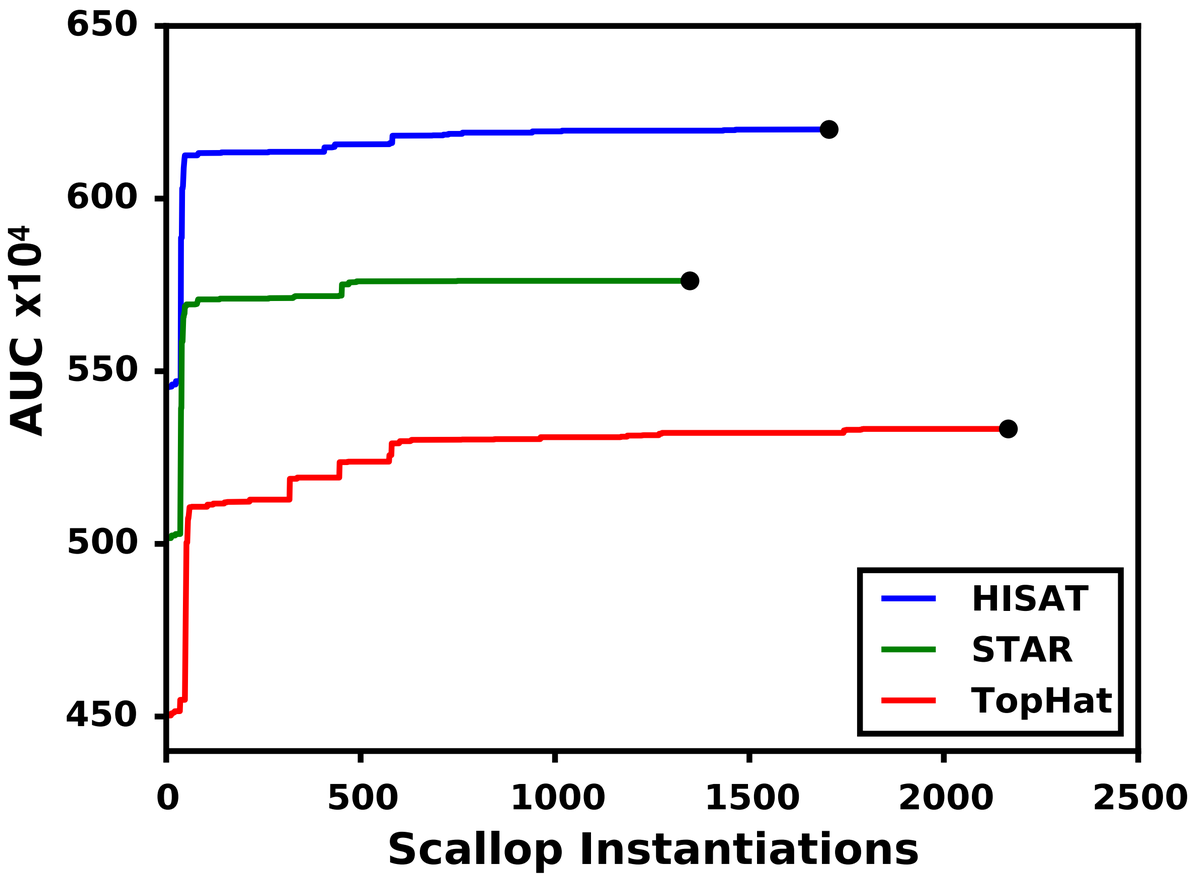


[DKK, WCB@ICML 2019]

Finding an advisor set

Parameter curve smoothness and single maxima help parameter selection.

- Iterative optimization will work well.
- Process is slow.



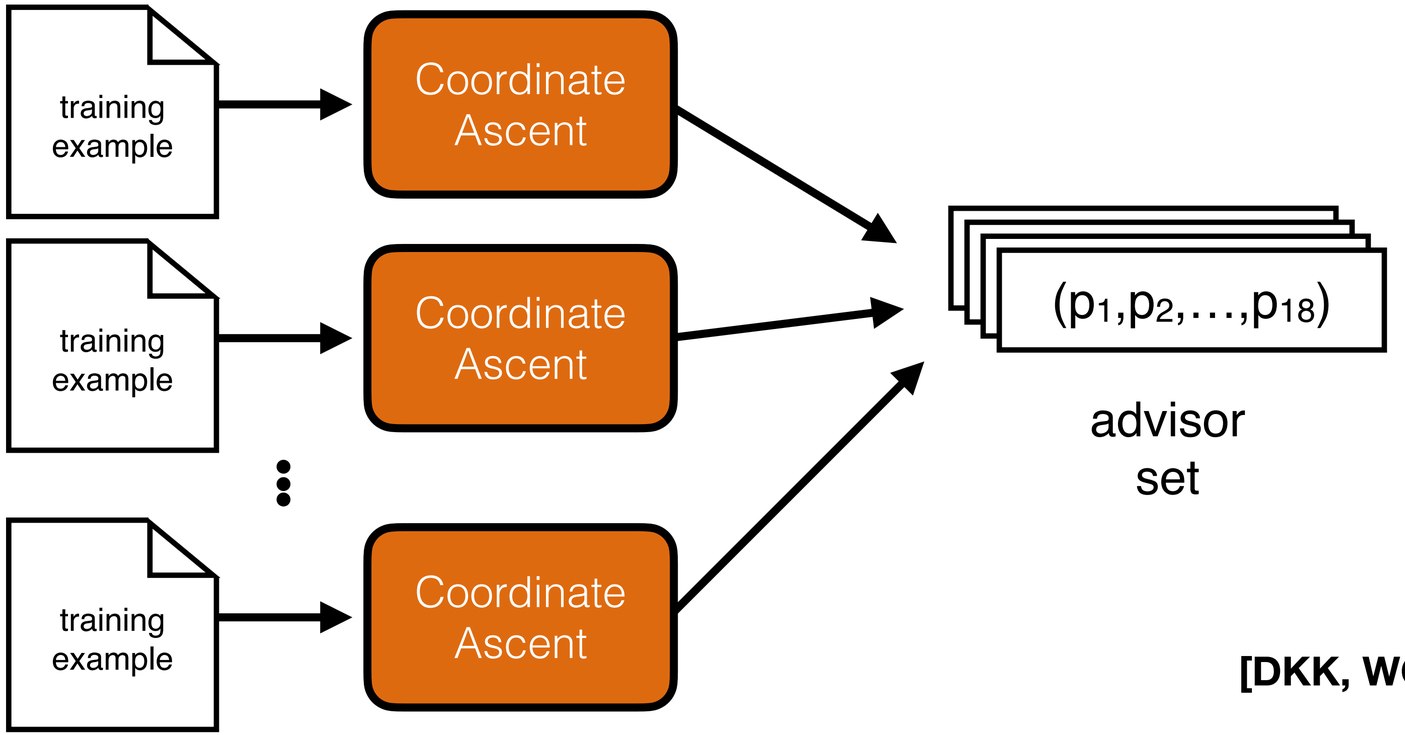
[DKK, WCB@ICML 2019]

[HISAT: Kim, et al. Nat. Met. 2015] [STAR: Dobin , et al., Bioinformatics 2013] [TopHat: Kim, et al., Gen. Bio. 2013]

Finding an advisor set

We can use coordinate ascent to find optimal parameter vectors.

- Training samples should cover the range of expected input.
- Settings are found for all 18 tunable parameters.
- Collection of produced vectors is advisor set.

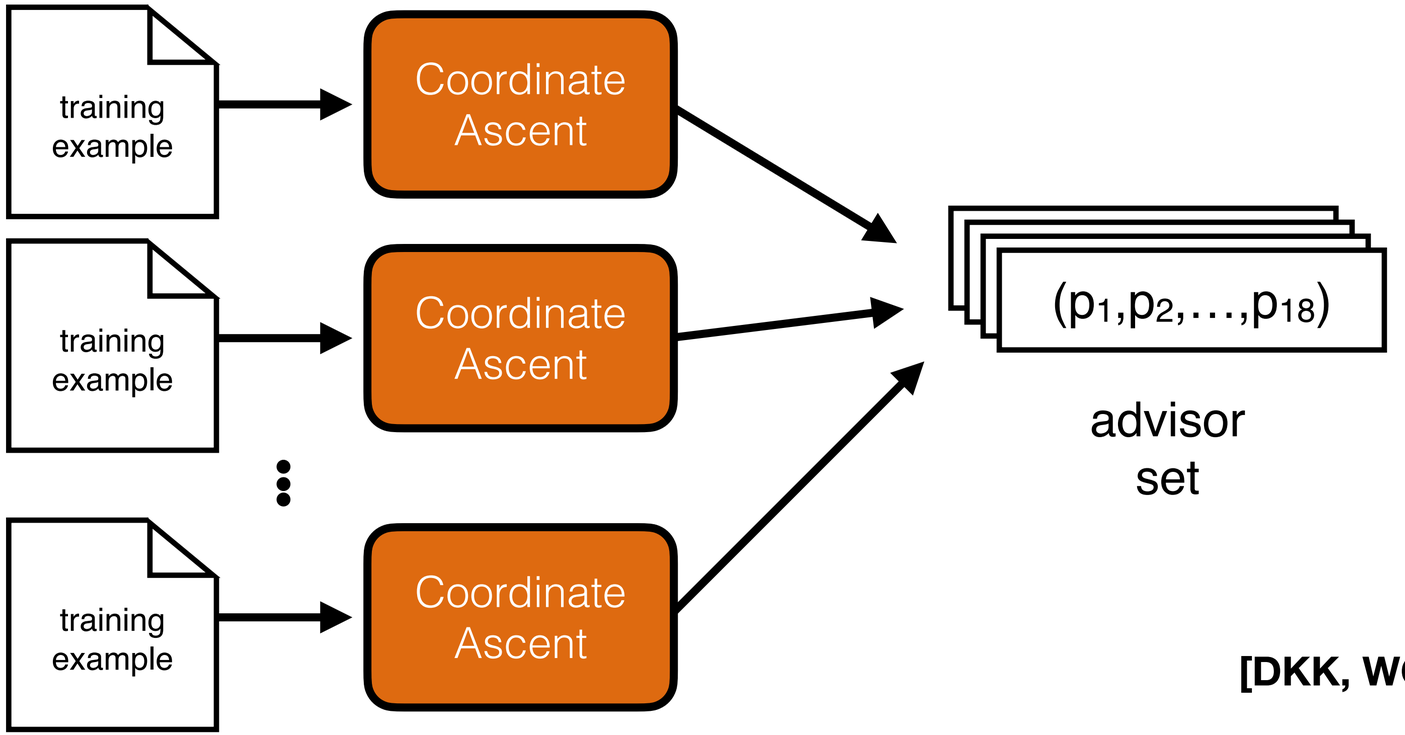


[DKK, WCB@ICML 2019]

Finding an advisor set

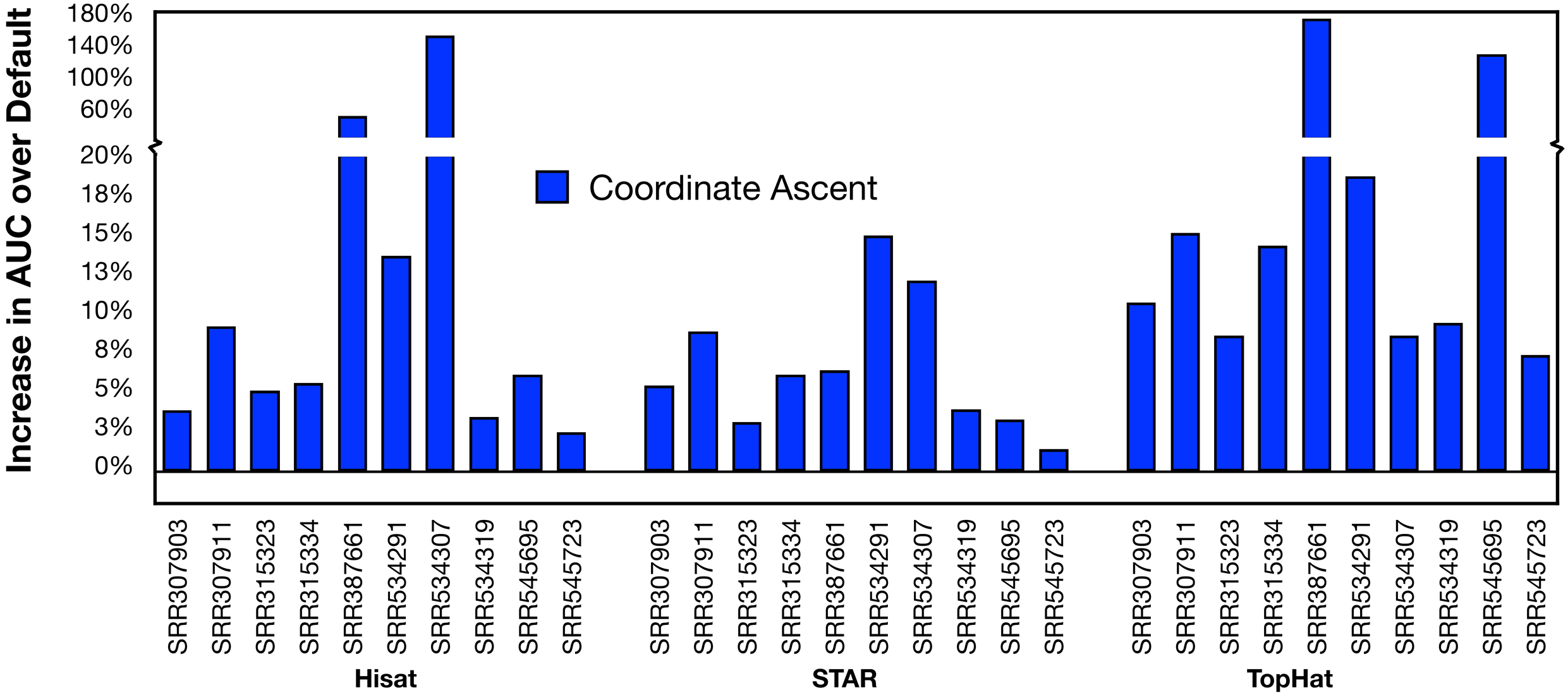
We can use coordinate ascent to find optimal parameter vectors.

- Training samples should cover the range of expected input.
- Settings are found for all 18 tunable parameters.
- Collection of produced vectors is advisor set.
- The set is precomputed and doesn't impact the advising time.



[DKK, WCB@ICML 2019]

Scallop advising

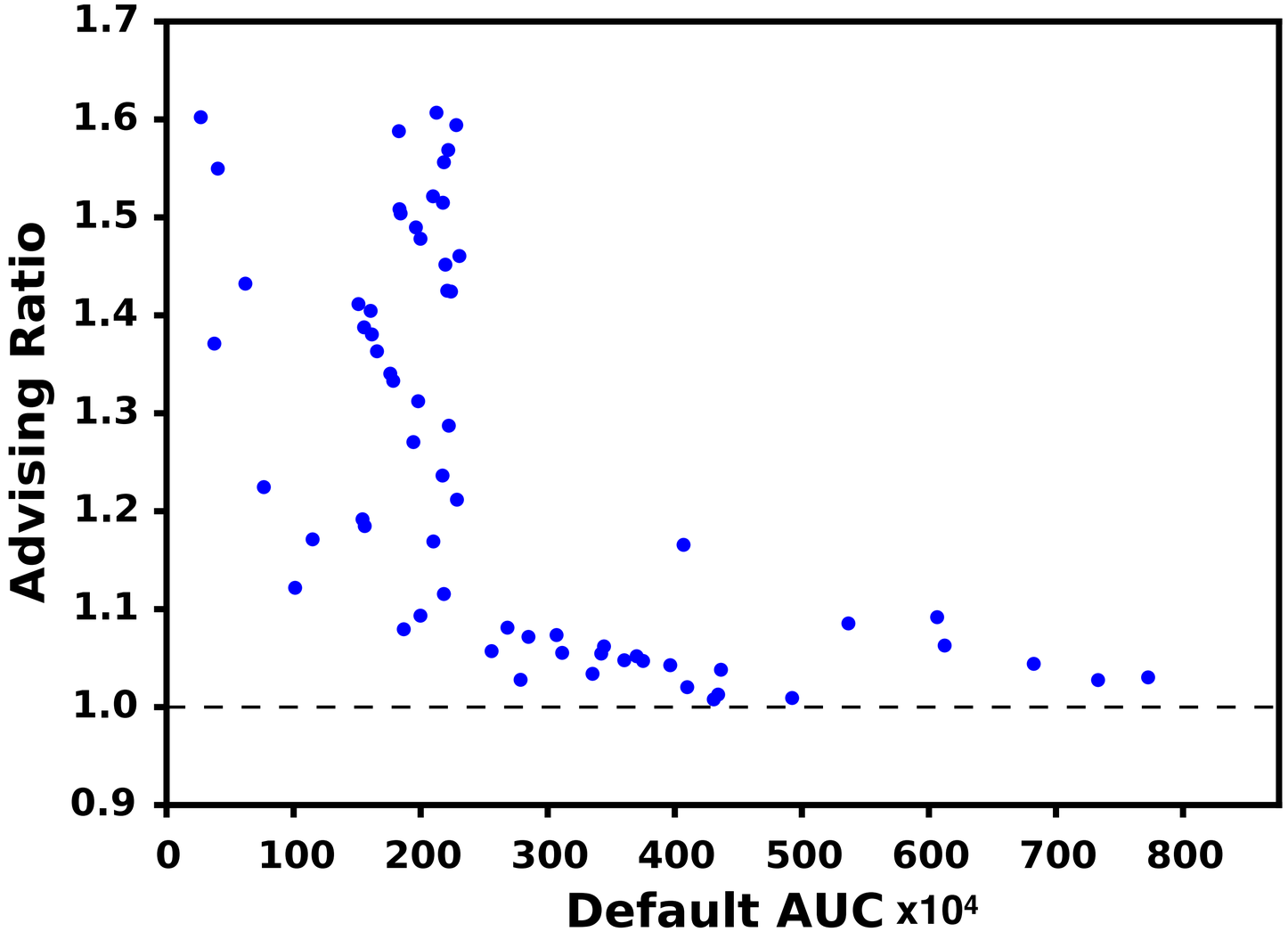


Average of 18.1% increase in AUC using Coordinate Ascent

[DKK, WCB@ICML 2019]

Scallop advising

- all aligned RNA-seq from ENCODE
- variety of aligners
- example of performance in general

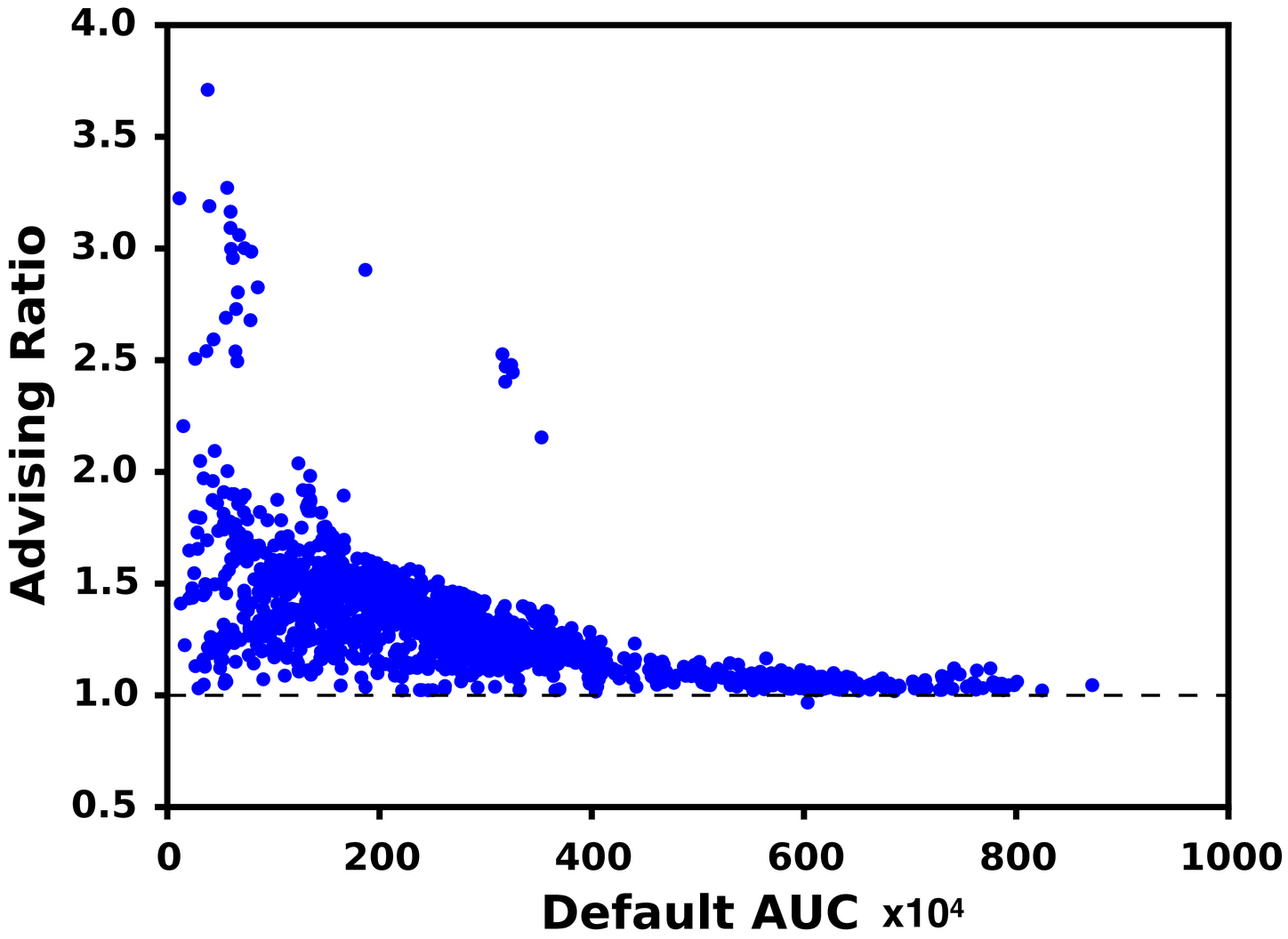


average advising ratio: 1.257

[DKK, WCB@ICML 2019]

Scallop advising

- 1595 RNA-Seq from SRA
- aligned using STAR
- example of high-throughput performance



average advising ratio: 1.382

[DKK, WCB@ICML 2019]

Genome Assembly

The first step in many genomic analyses is to map the reads from the individual to a reference genome.

Once the reads are mapped, we can identify the changes between the input and the reference.

Many tools exist to perform this task, each with:

- a large number of tunable parameters, and
- different performance characteristics

Unlike transcript assembly, no ground truth,

- unless we use simulation.

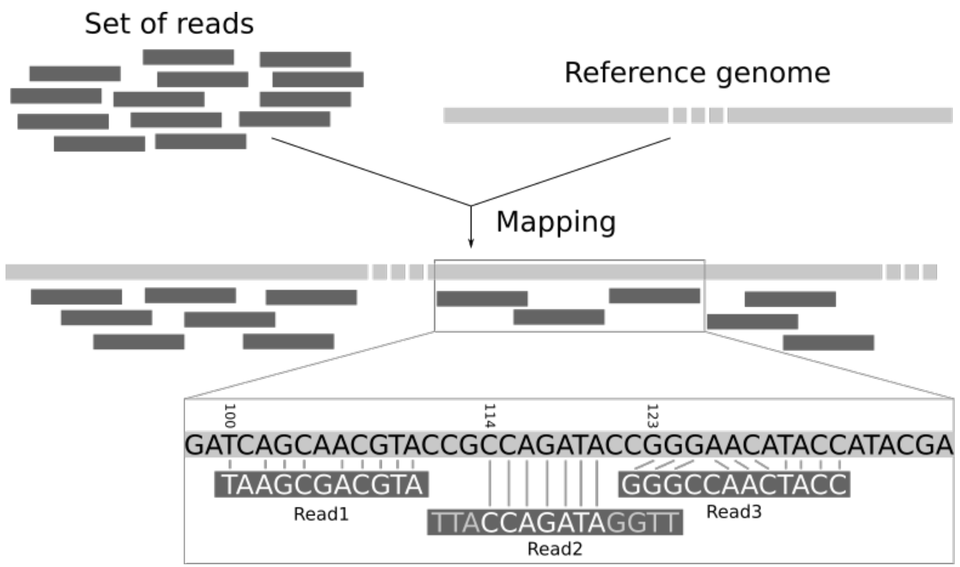


Image courtesy of The Galaxy Project:

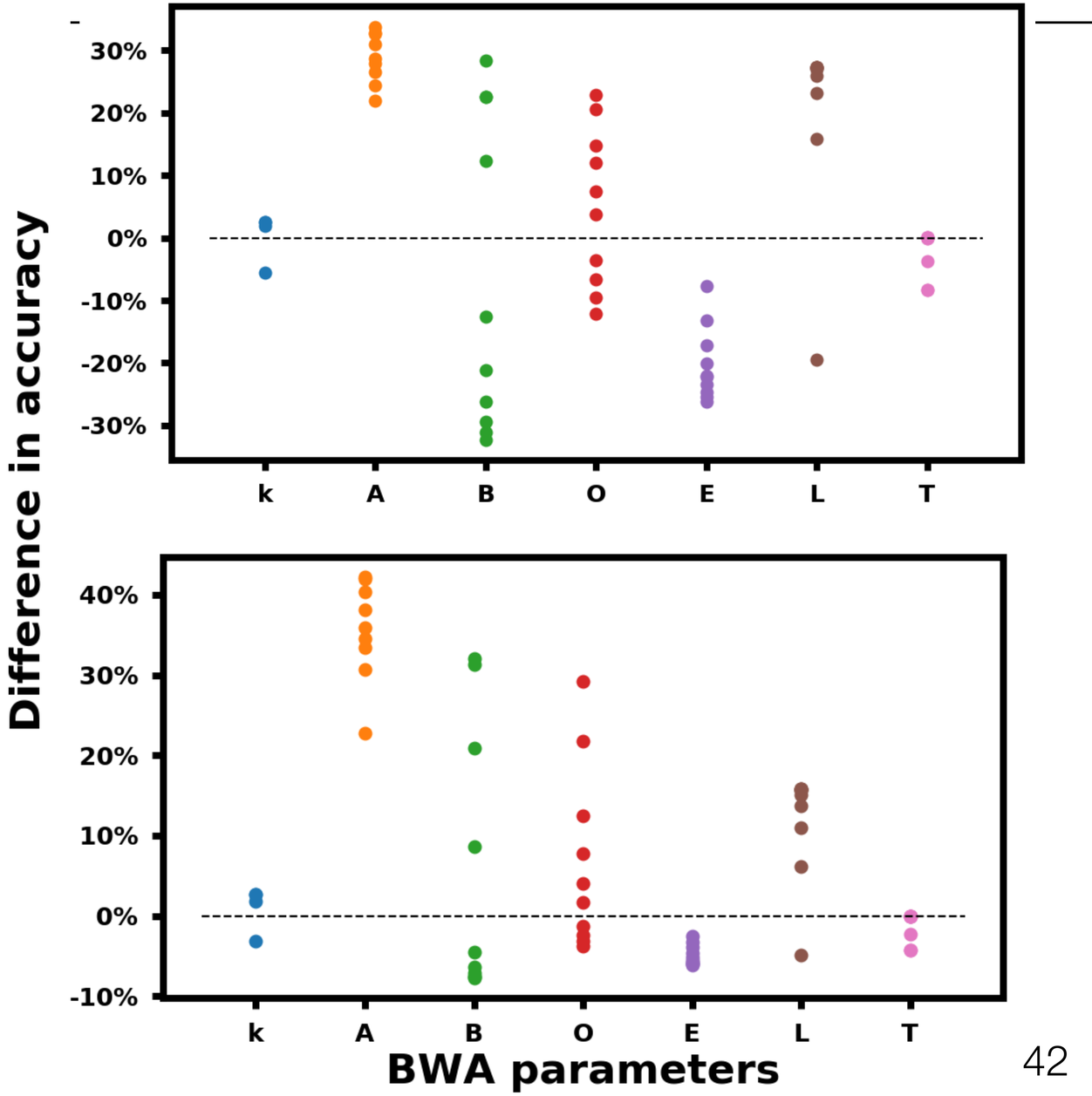
https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Genome Assembly

Two simulated datasets generated using different simulation parameters.

Parameter vectors with only one parameter value changed away from its default.

BWA *can* be improved significantly, but only if the parameter choice changes are selected carefully



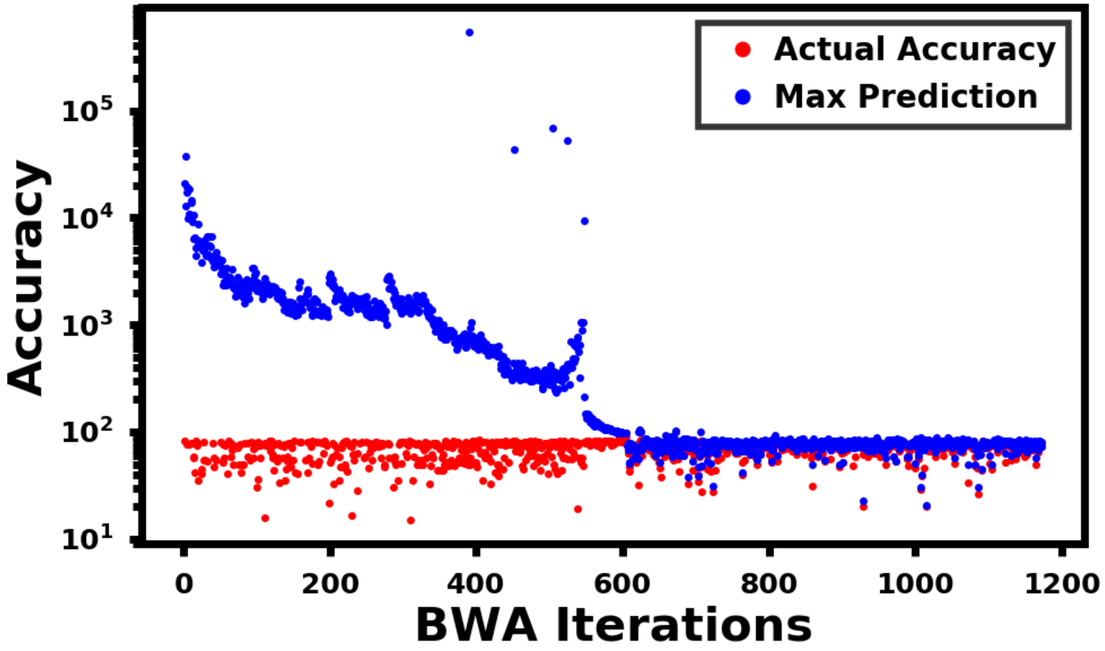
Genome Assembly

To explore the parameter space we

- use a polynomial accuracy estimator
- with the parameters of BWA as the input features,
- and an active learning approach.

Each vertical position is one learning instance

- find the highest predicted accuracy parameter vector,
- run assembly and add to training,
- repeat until prediction is correct.



Minimizer Schemes for Genome Analysis

Sequence Similarity

Sequence similarity is used in many contexts:

- comparing web pages
- suggestion systems
- finding plagiarism
- matching sequencing reads
- binning genetic material

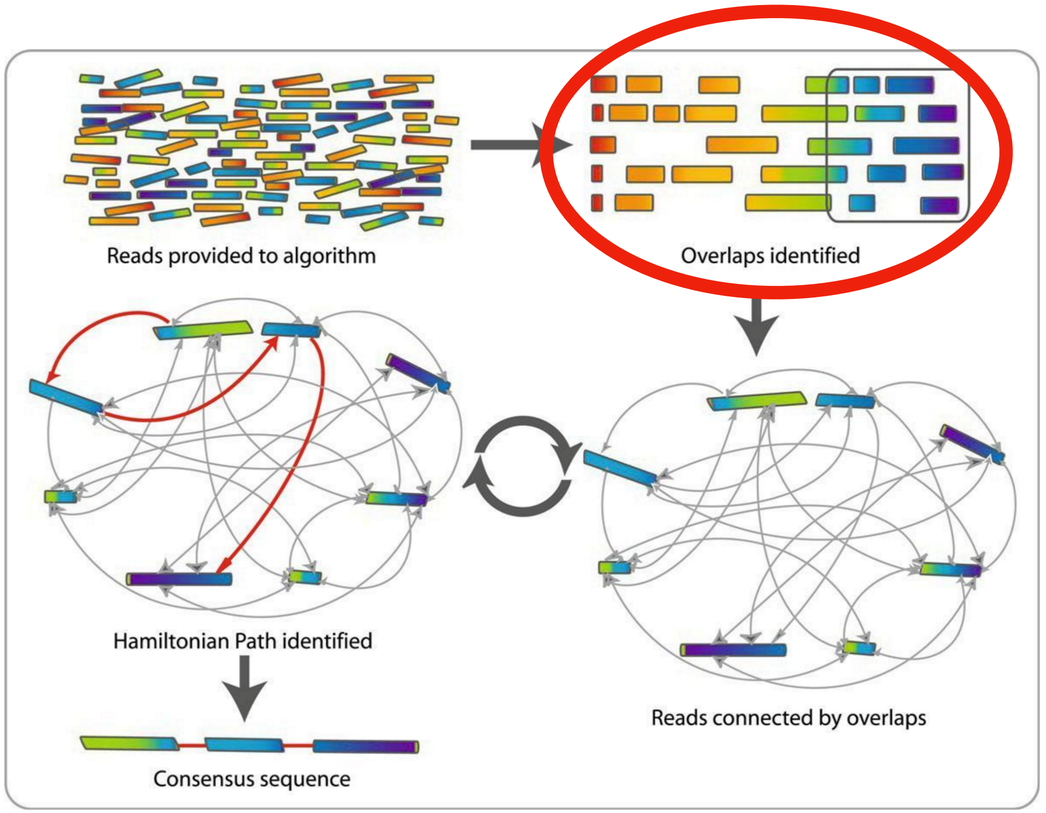
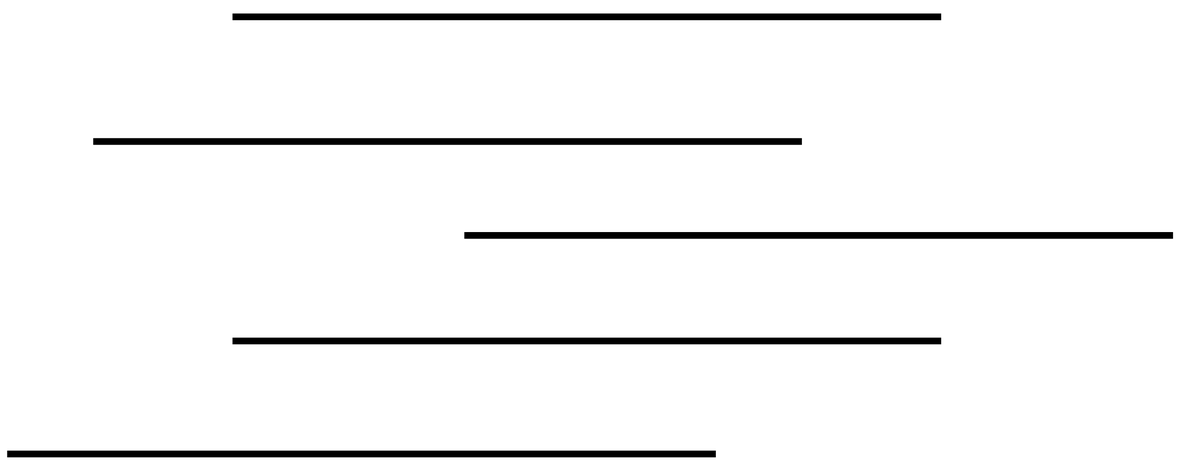


Image courtesy of The Galaxy Project:
https://training.galaxyproject.org/archive/2021-08-01/topics/assembly/images/olc_pic.png

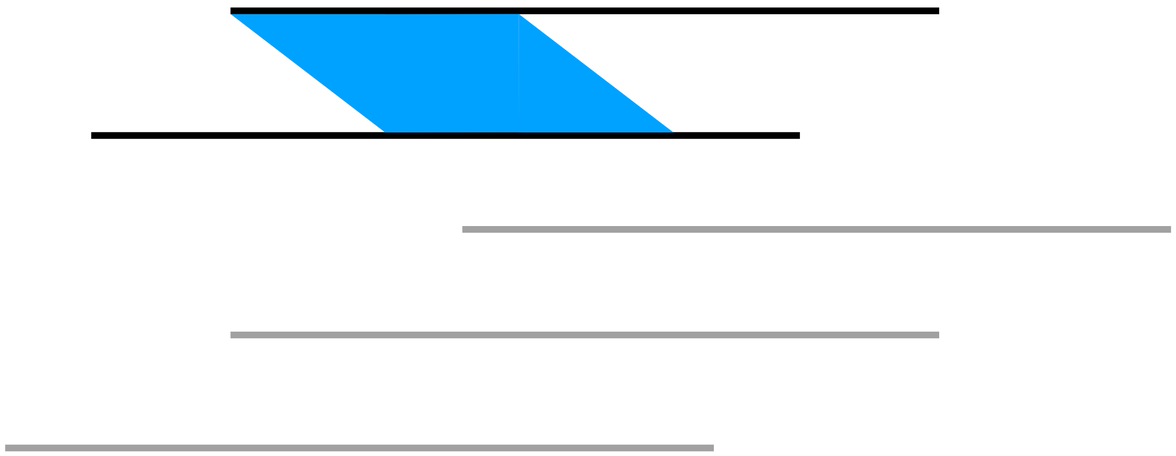
Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



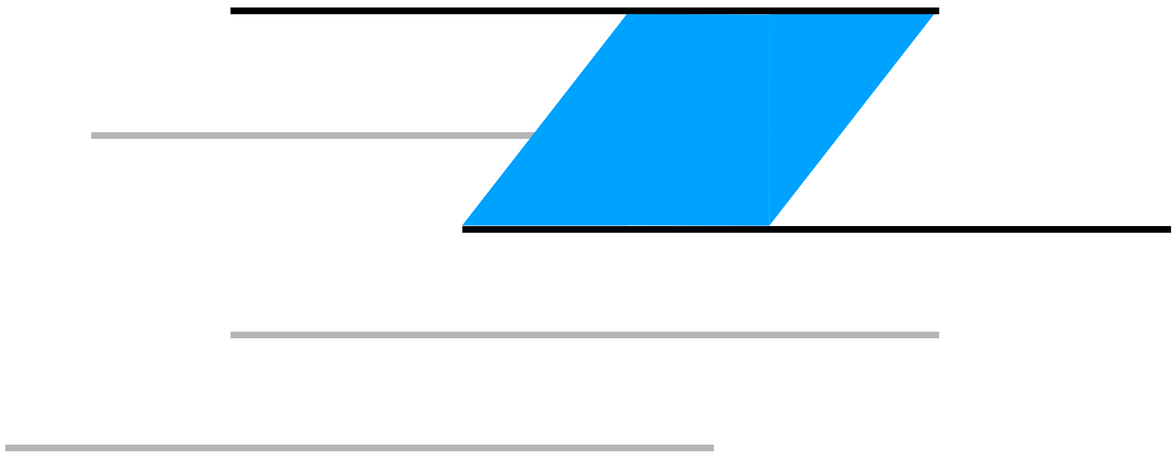
Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



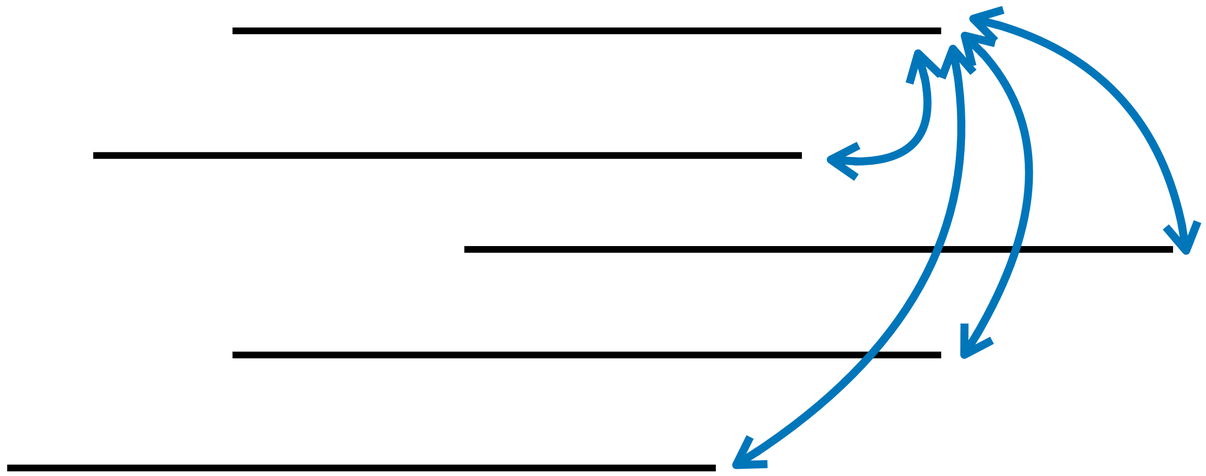
Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



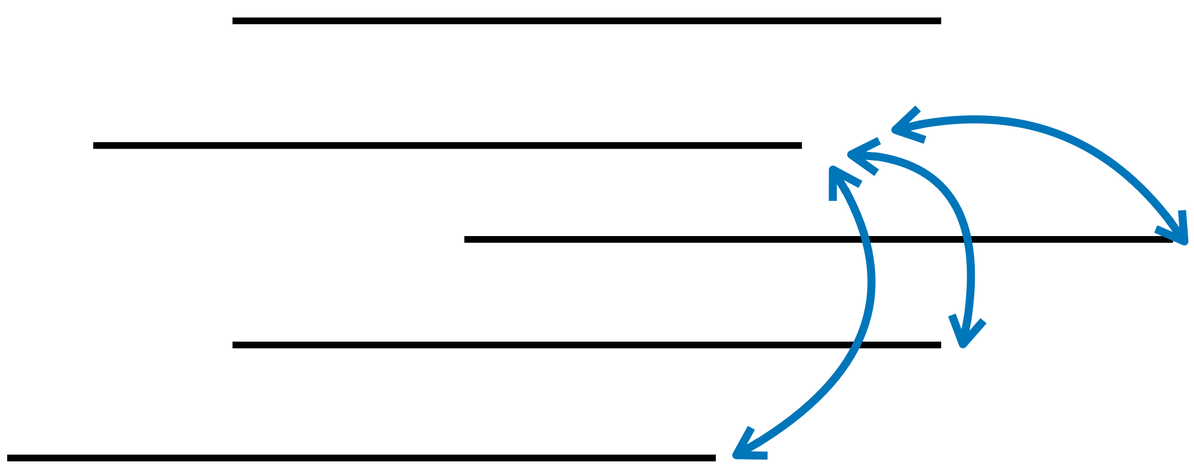
Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



Minimizer Schemes

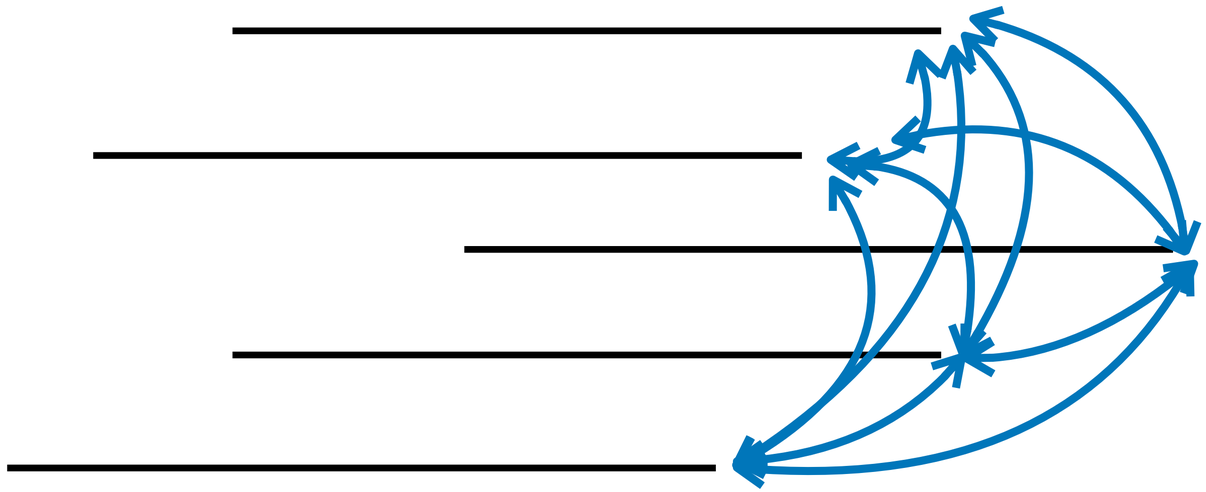
Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



Minimizer Schemes

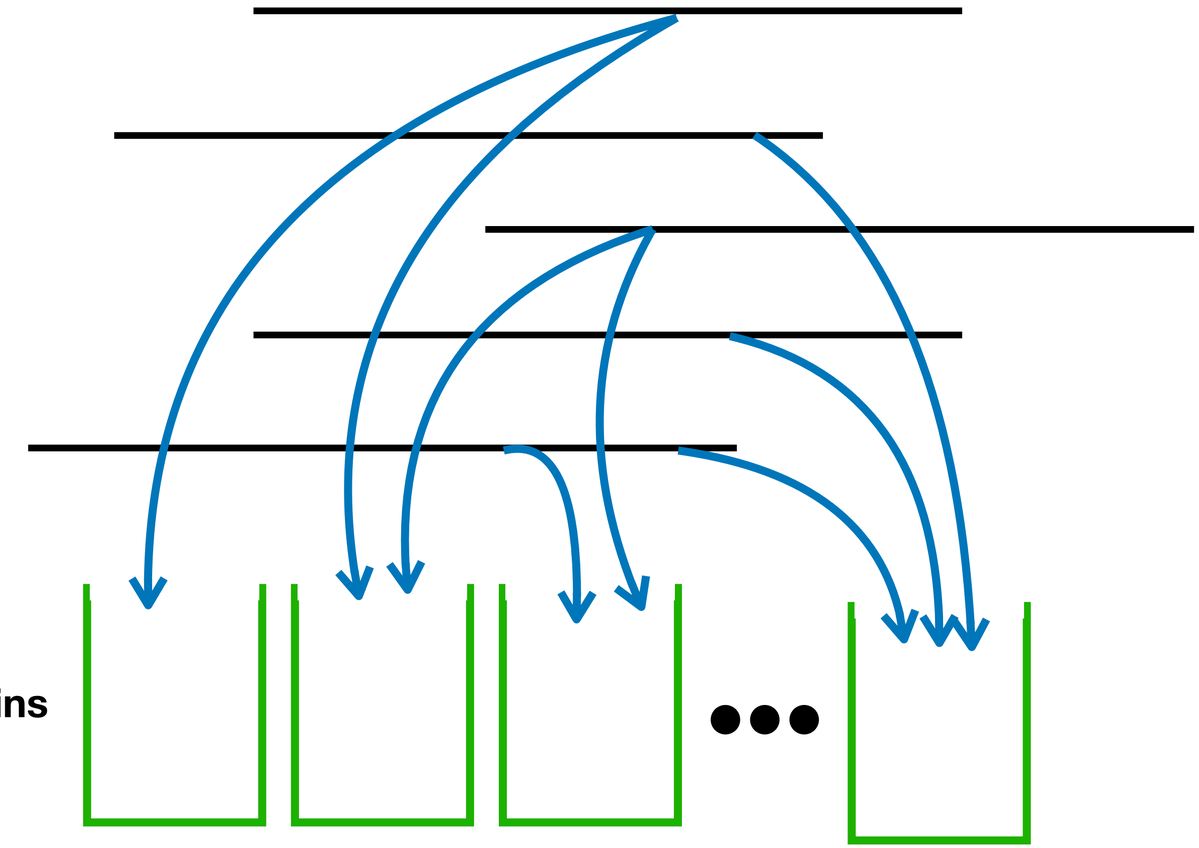
Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation

$O(n^2)$ alignments!



Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



Only compare within bins

Minimizer Schemes

Minimizer schemes have two special properties:

- two sequences with a long exact match must select the same *k*-mers
- there are no large gap between selected *k*-mers

Minimizer Schemes

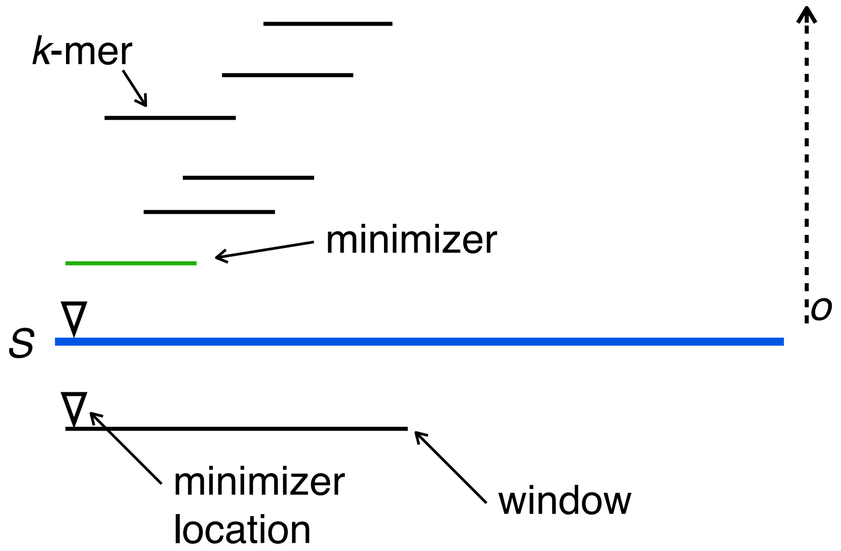
Minimizer schemes have two special properties:

- two sequences with a long exact match must select the same *k*-mers
- there are no large gap between selected *k*-mers

Use in *k*-mer counting, *de Brujin* graph construction, data structure sparsification, etc.

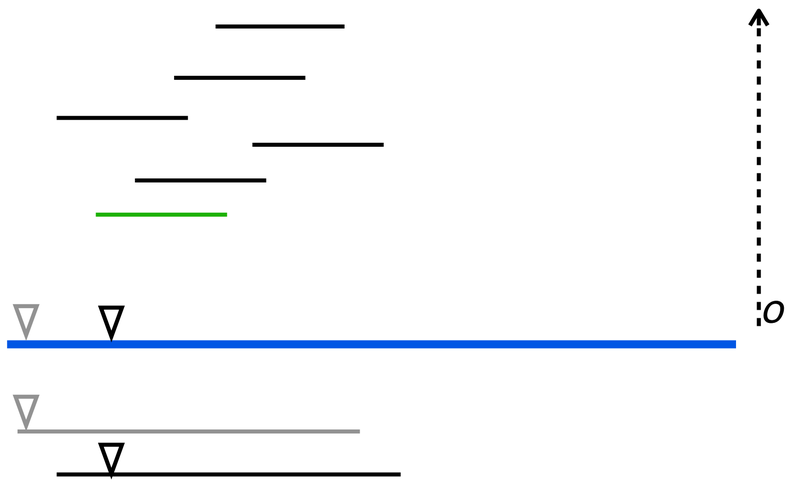
Minimizer Schemes

For a windows of *w* consecutive *k*-mers from a sequence *S*, a minimizer scheme selects the minimum according to an ordering o as a representative



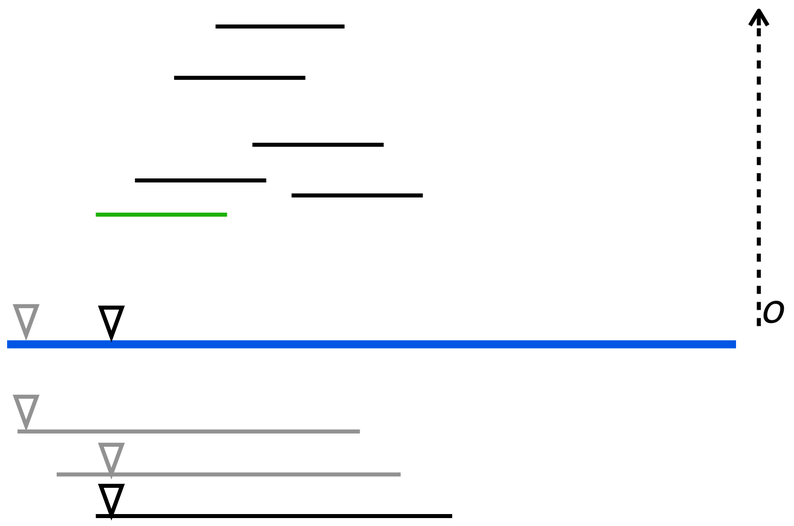
Minimizer Schemes

For a windows of *w* consecutive *k*-mers from a sequence *S*, a minimizer scheme selects the minimum according to an ordering o as a representative



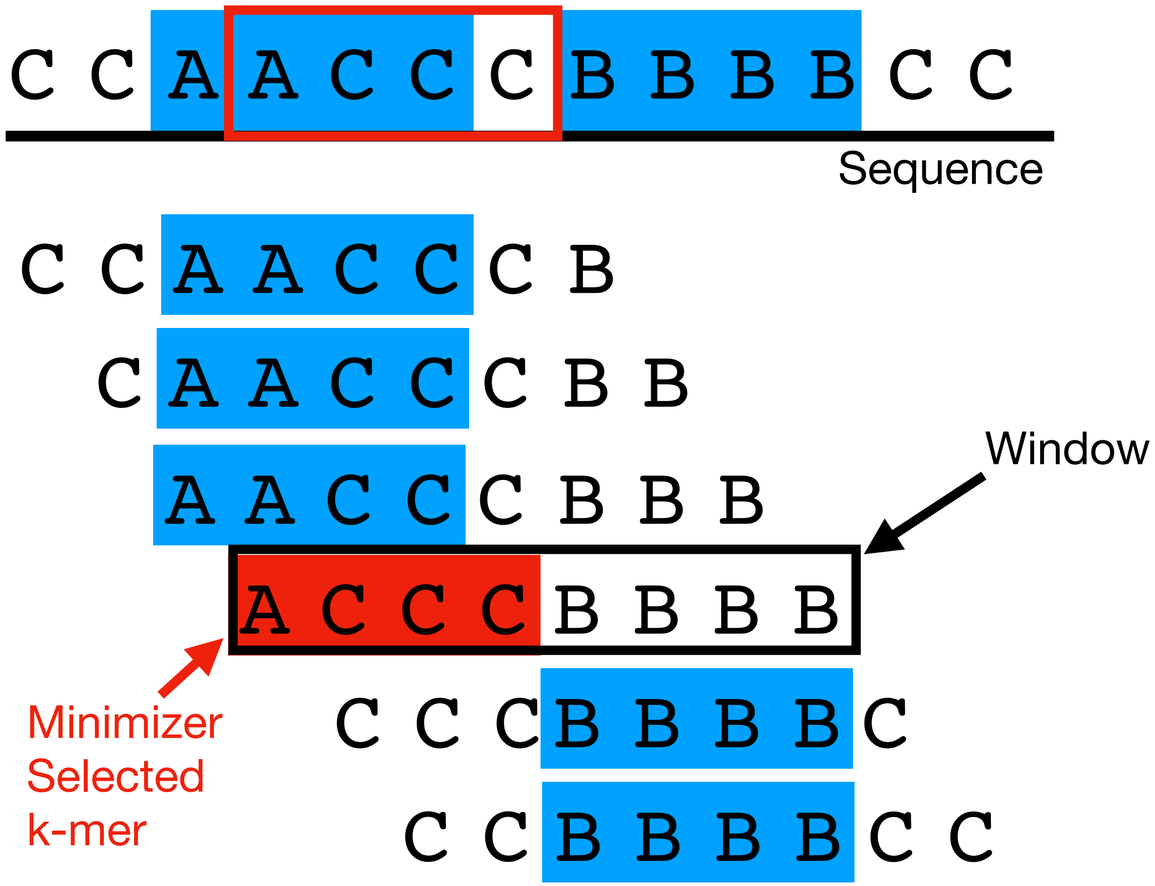
Minimizer Schemes

For a windows of *w* consecutive *k*-mers from a sequence *S*, a minimizer scheme selects the minimum according to an ordering o as a representative



Minimizer Schemes

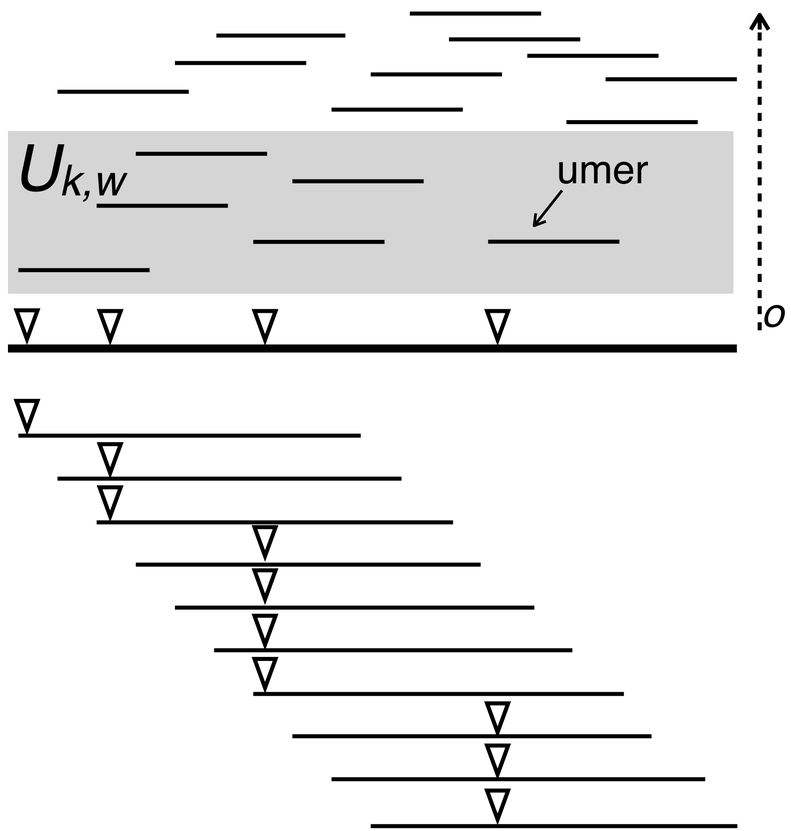
An extra example



Universal *k*-mer Set and Minimizer Ordering

A **universal *k*-mer set** induces a family of compatible minimizer orderings

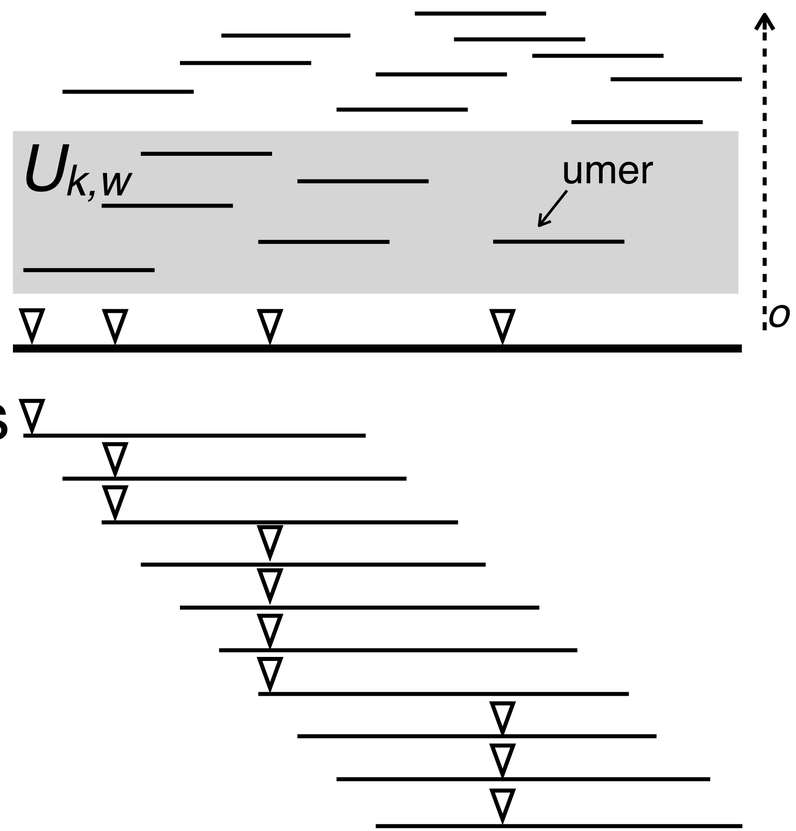
- A universal *k*-mer set $U_{k,w} \subseteq \Sigma^k$ is a set of *k*-mers such that any window of *w* consecutive *k*-mers must contain at least one element from the set



Universal *k*-mer Set and Minimizer Ordering

A **universal *k*-mer set** induces a family of compatible minimizer orderings

- A universal *k*-mer set $U_{k,w} \subseteq \Sigma^k$ is a set of *k*-mers such that any window of *w* consecutive *k*-mers must contain at least one element from the set

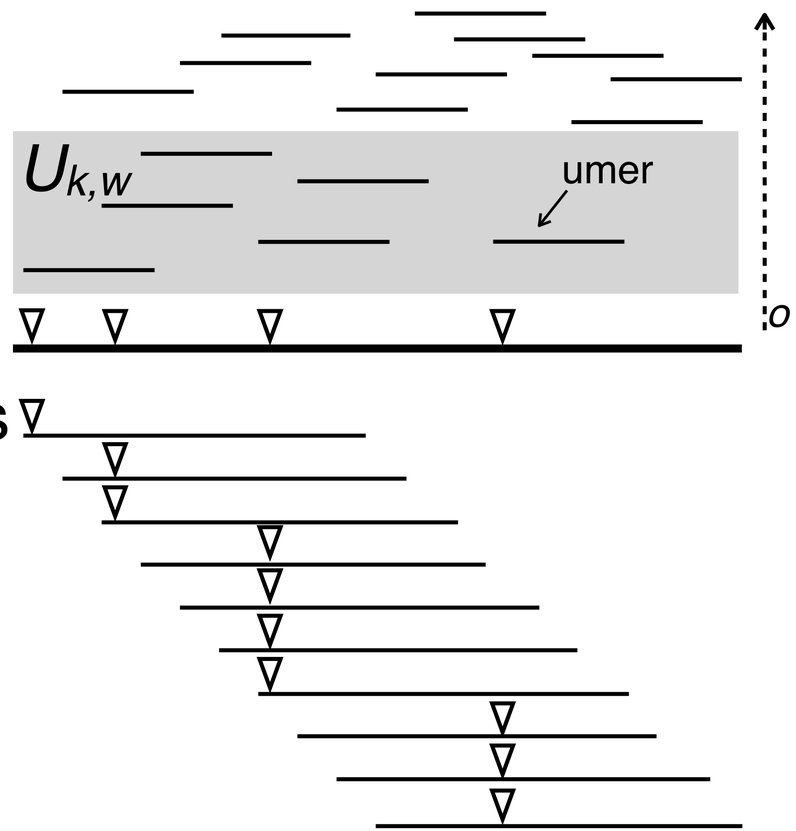


Orderings based on universal sets have better performance then lexicographic or random orders [Marçais, et al. 2017]

Universal *k*-mer Set and Minimizer Ordering

A **universal *k*-mer set** induces a family of compatible minimizer orderings

- A universal *k*-mer set $U_{k,w} \subseteq \Sigma^k$ is a set of *k*-mers such that any window of *w* consecutive *k*-mers must contain at least one element from the set

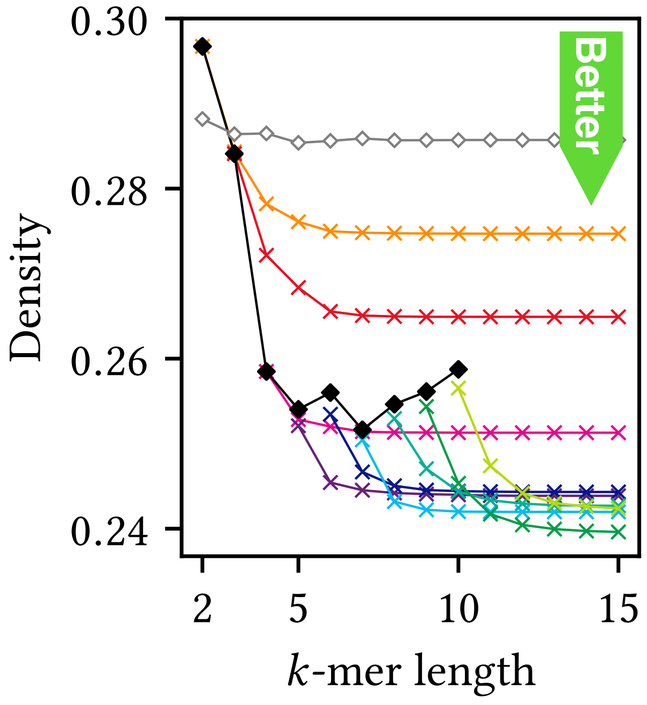


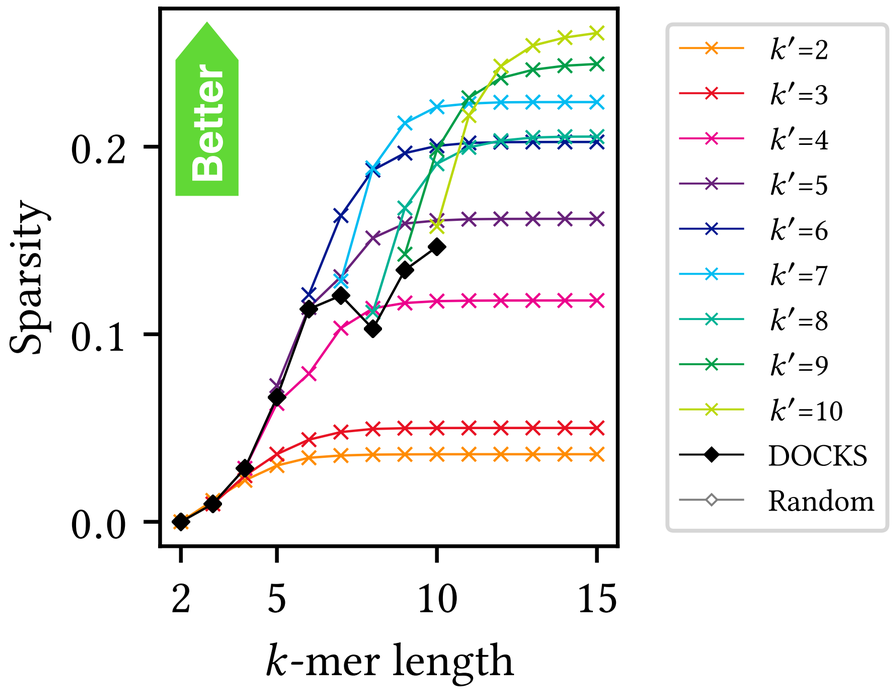
Orderings based on universal sets have better performance then lexicographic or random orders [Marçais, et al. 2017]

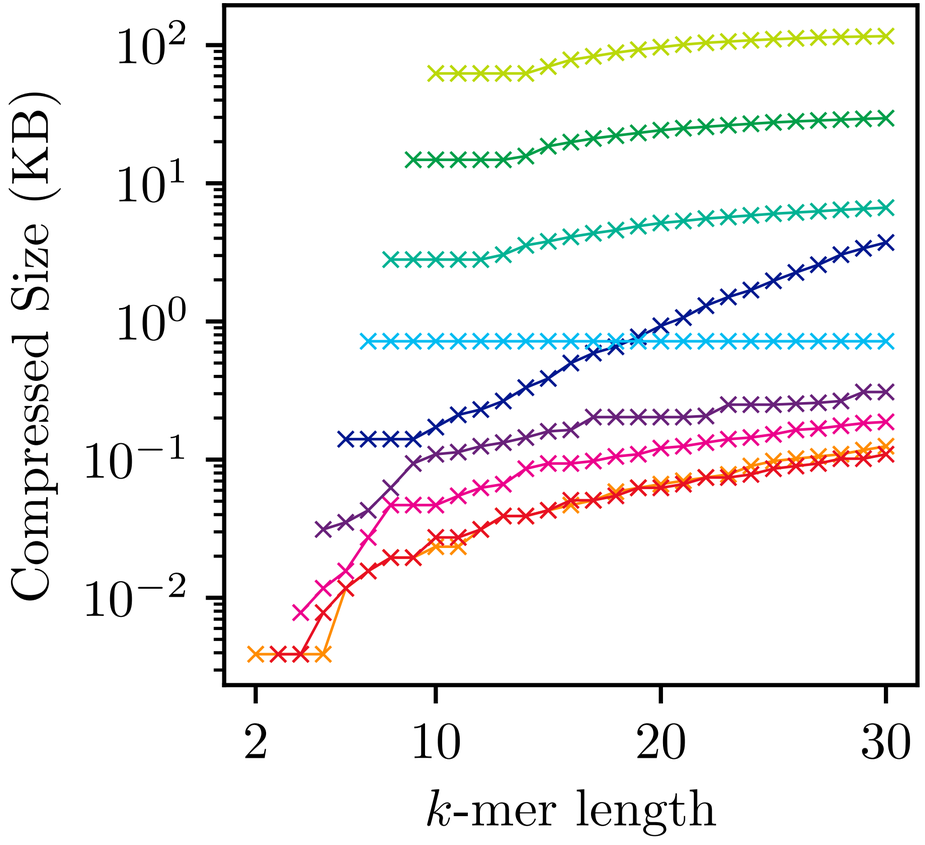
Recent work has shown that we can build universal sets for large *k* & *w* (like those used in practice) from existing sets for small *k* & *w* [DeBlasio, et al. 2019; Zheng, et al. 2020]

Storing a universal set is inefficient

Stored using a sequence trie, high complexity leads to large files

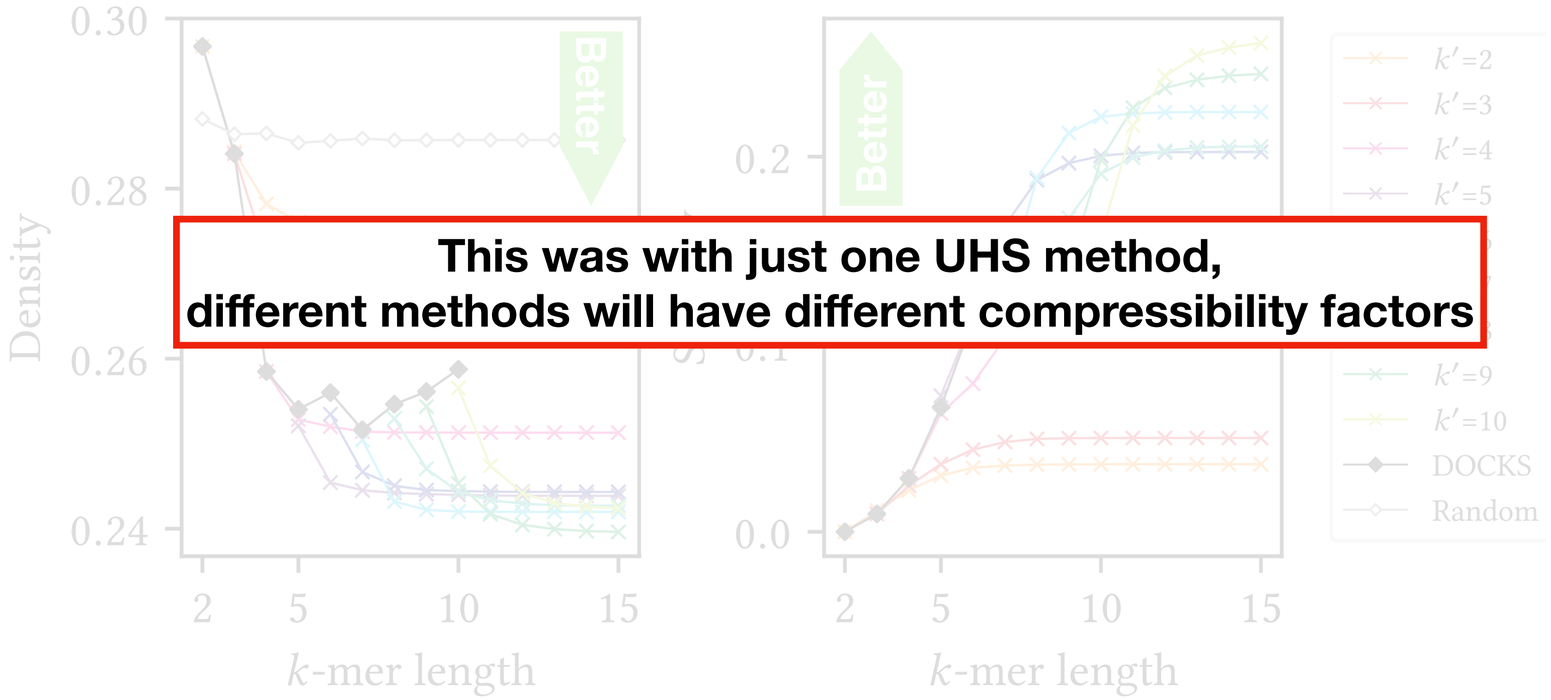




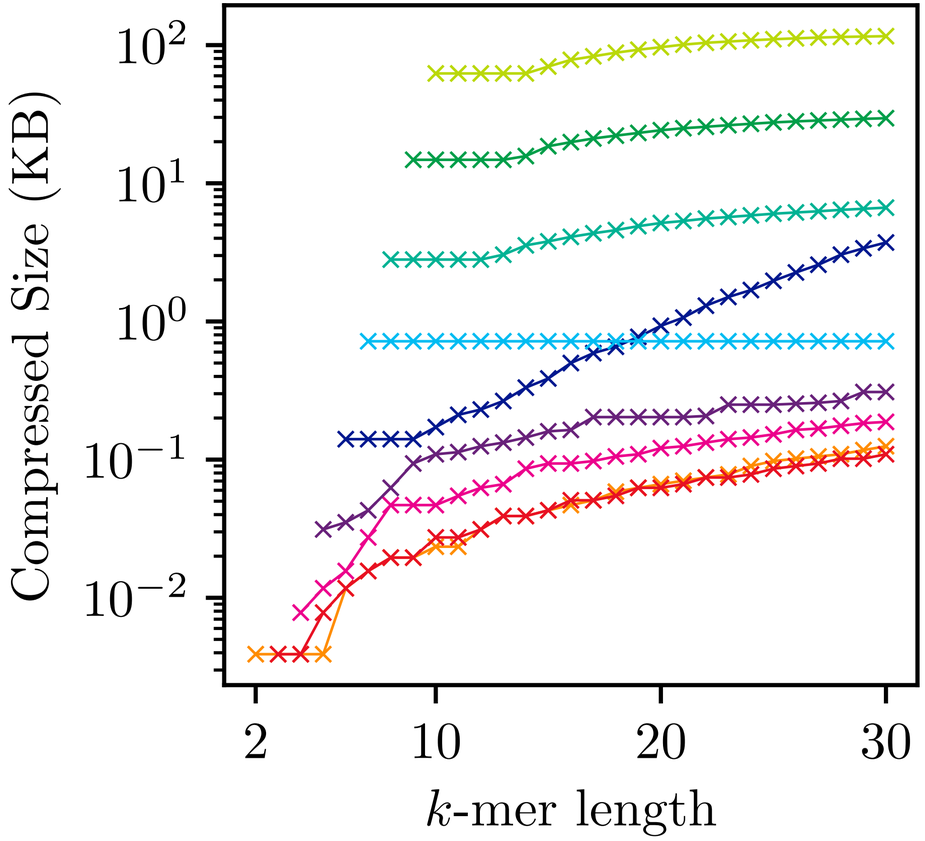


Storing a universal set is inefficient

Stored using a sequence trie, high complexity leads to large files



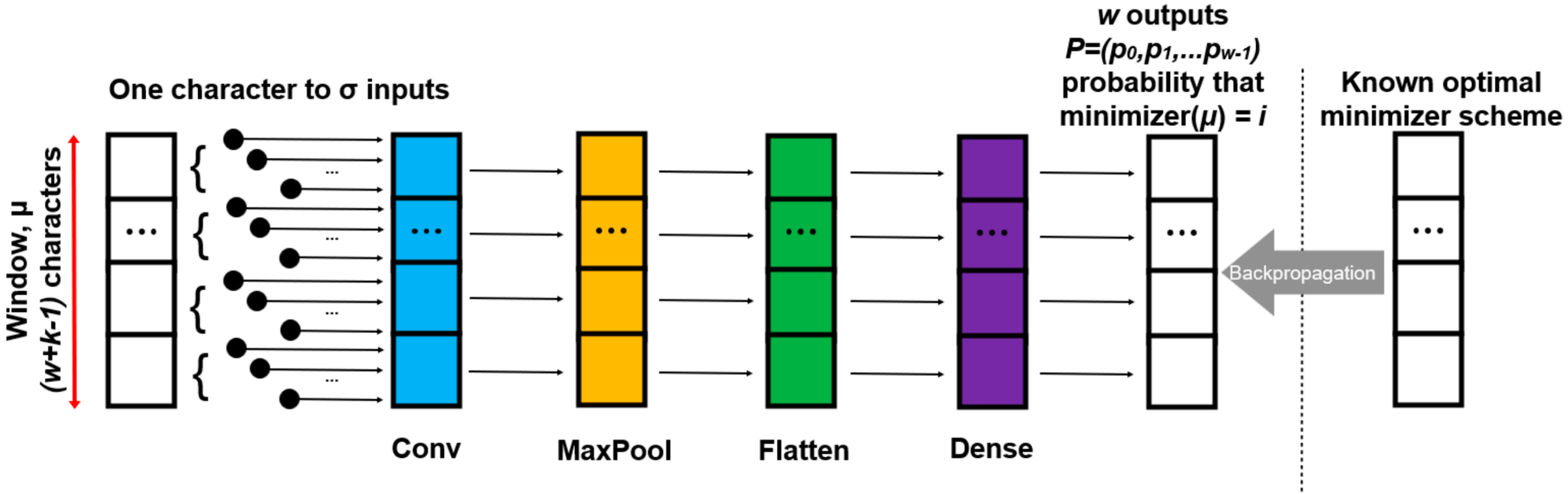
This was with just one UHS method,
different methods will have different compressibility factors



Our proposed method

Task -- learn the minimizer schemes using back propagation

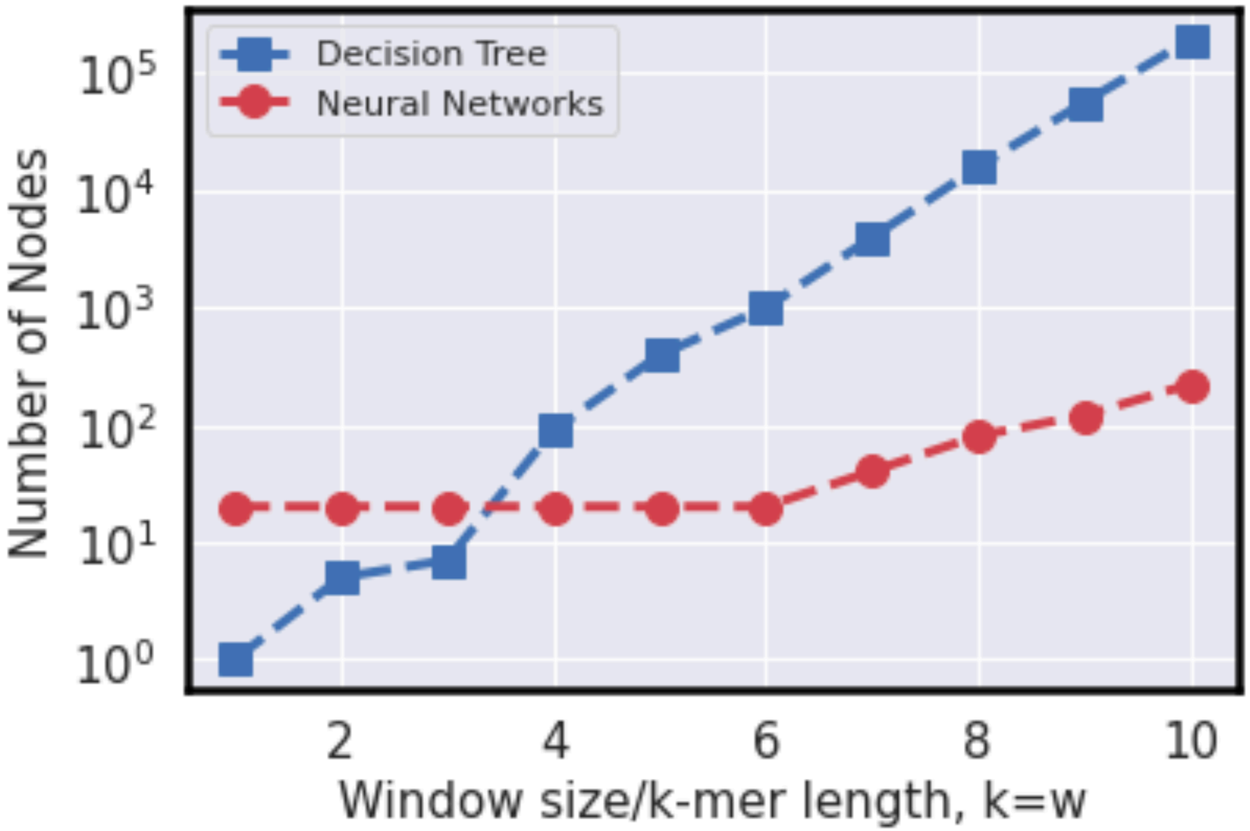
- Our task is to create a network topology is complex enough to encode existing schemes, but not so complicated that it provides extreme training times.
- One issue that arises is that for small values of w and k there may not be enough information to train the network completely since there are only so many unique windows.



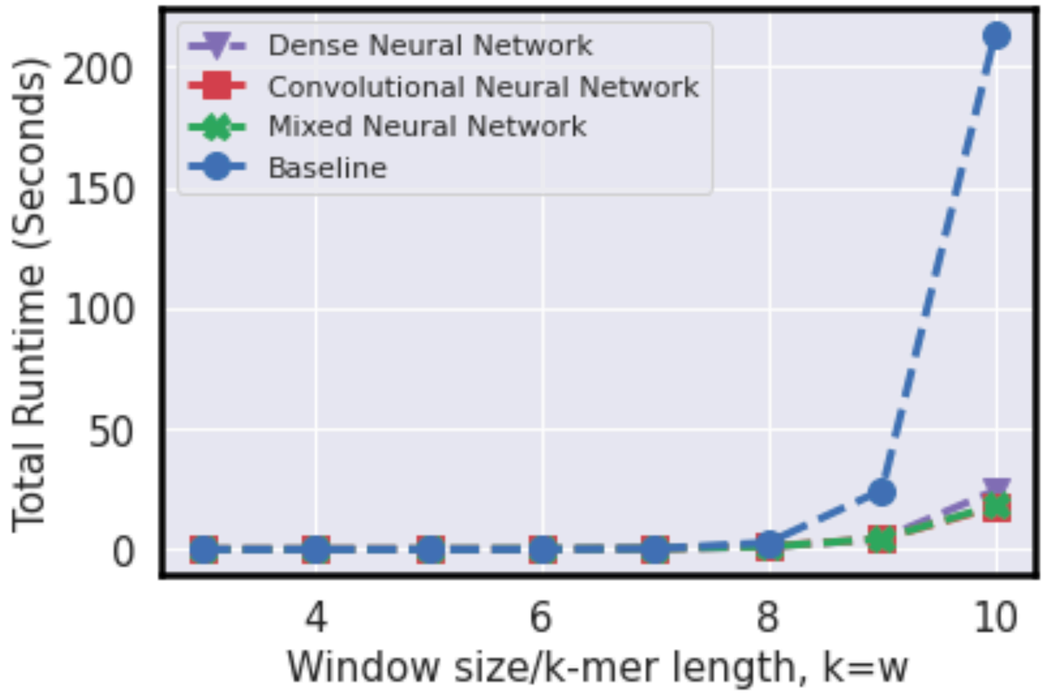
A note about Neural Networks

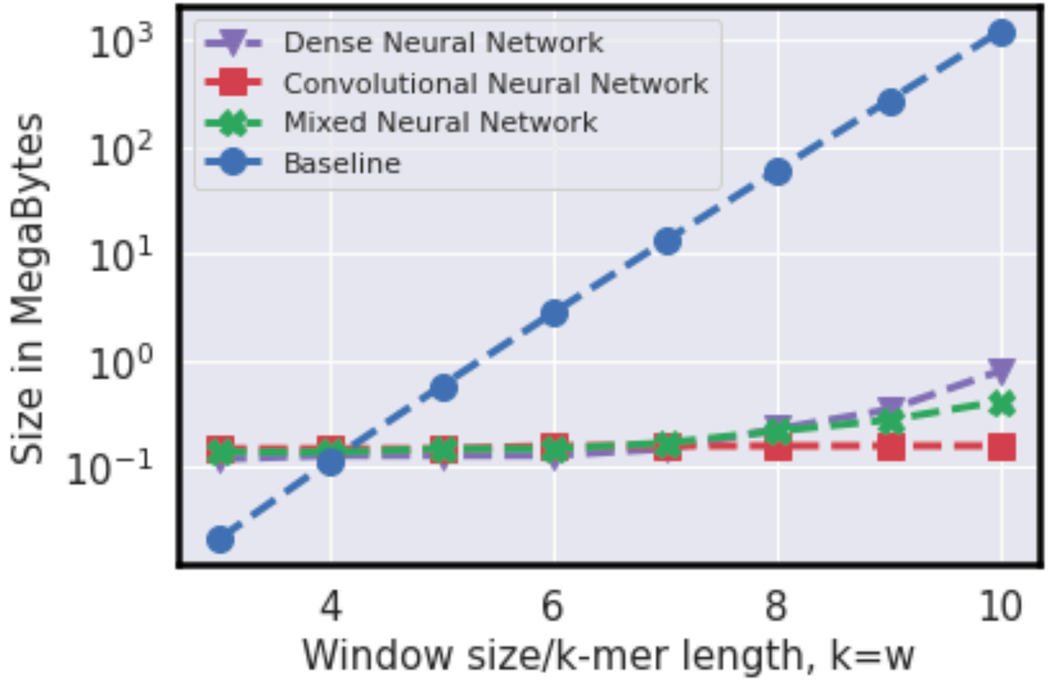
Used Decision Trees and Dense Neural Networks.

The number of nodes to encode minimizers is significantly larger with decision trees than with neural network implementations.



Performance of the networks





A trained model has a shorter k-mer lookup time and smaller memory footprint than a naïve implementation of minimizers.

Neural Networks for Object Identification

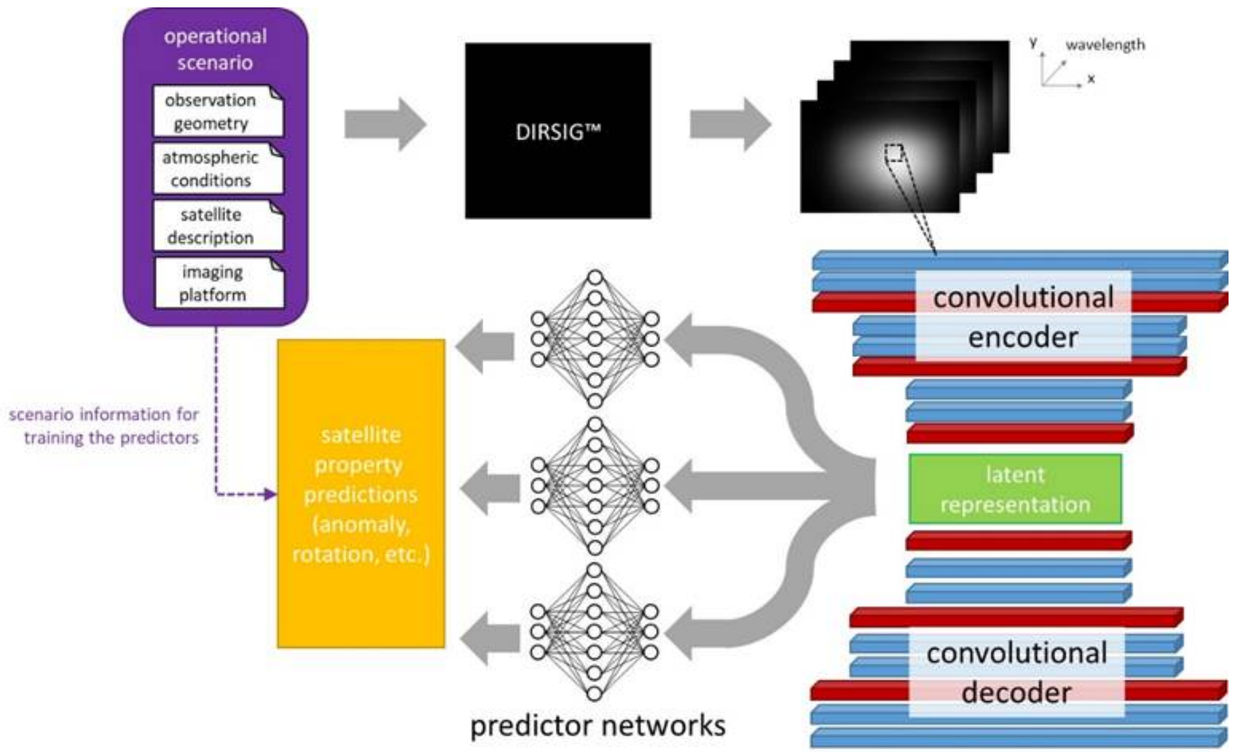
Identifying attributes of satellites is non-trivial

Hyperspectral (HSI) and polarization imaging systems are becoming available and provide geographical spatial diversity.

Accurate interpretation of these images may allow us to perceive, predict, comprehend, and react appropriately to changing situations in the space domain.

HSI ground-based observation systems collect spectro-temporal signatures of Unresolved Resident Space Objects (URSOs).

The high-spectral resolution allows for the extraction of properties/parameters of the URSO using spectral domain information even though it cannot be resolved in the spatial domain.



Acknowledgments

The DeBlasio Lab

Luis Cedillo (UG, Facet-NN/LR)
Demetrius Hernandez (UG, Minimizers)
Taposh Sarkar (PhD, USROs)
Fernando Sepulveda (UG)
Md. Easin Hasan (former, BWA)
Hector Richart-Ruiz (former)

The Current Collaborators

Miguel Velez-Reyes
Arizbe Najera

Contact

deblasiolab.org
danfdeblasio

Previous Collaborators

John Kececioglu
Travis Wheeler (Montana)
Jen Wisecaver (Purdue)

Carl Kingsford
Fiyinfoluwa Gbosibo
Kwanho Kim
Guillaume Marçais

Funding