Boosting alignment accuracy through adaptive local realignment

Dan DeBlasio

Computational Biology Department
Carnegie Mellon University*

John Kececioglu

Department of Computer Science
University of Arizona

Computational Biology Department | Carnegie Mellon University School of Computer Science

*work performed at
The University of Arizona

The University of Arizona | Computer Science UA Science

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

Motivation

Aligners often use *one* default parameter choice for *all* inputs.

- The default has good *average accuracy* across all benchmarks.
- The optimal default choice can be found by inverse alignment.

[Kececioglu and Kim 2007]

Motivation

Aligners often use *one* default parameter choice for *all* inputs.

- The default has good *average accuracy* across all benchmarks.
- The optimal default choice can be found by inverse alignment.
- The default may be a poor choice for specific inputs.

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 ...
```

Motivation

Proteins can have different mutation rates along their length

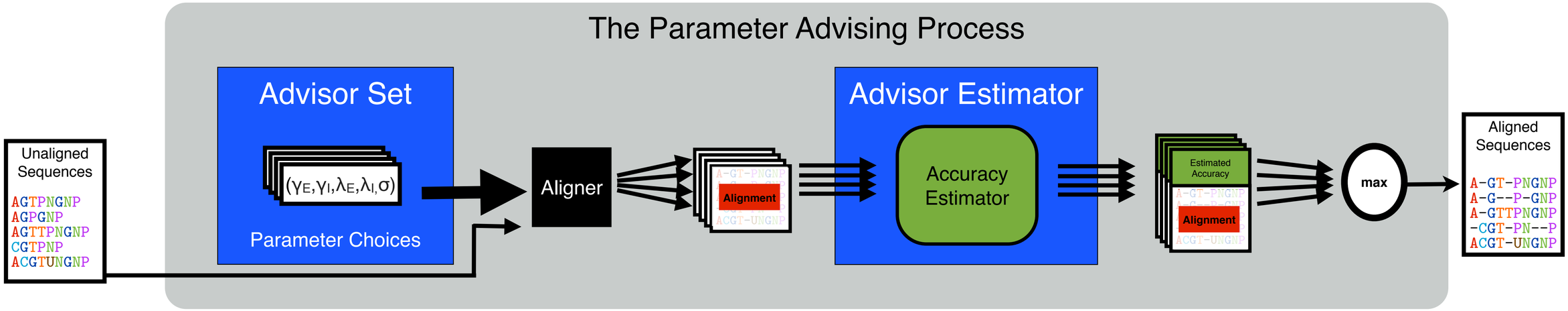
- Alignment parameters model mutation rates
- A single choice of parameters may not be best
- Using different choices across the protein can be superior

Can we find a parameter choice
that is best for *each region*
of a given input?

Parameter advising

Advising for unaligned input sequences

- aligns sequences using each parameter choice from a set,
- assigns an estimated accuracy to each alignment, and
- selects the alignment with the highest estimated accuracy.



[RECOMB 2012]

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 aligned residue pairs 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")

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

[RECOMB 2012]

Advisor sets

An advisor set should

- consist of complementary parameter settings,
- produce at least one good alignment for every input,
- be small, to reduce running time.

An oracle set is

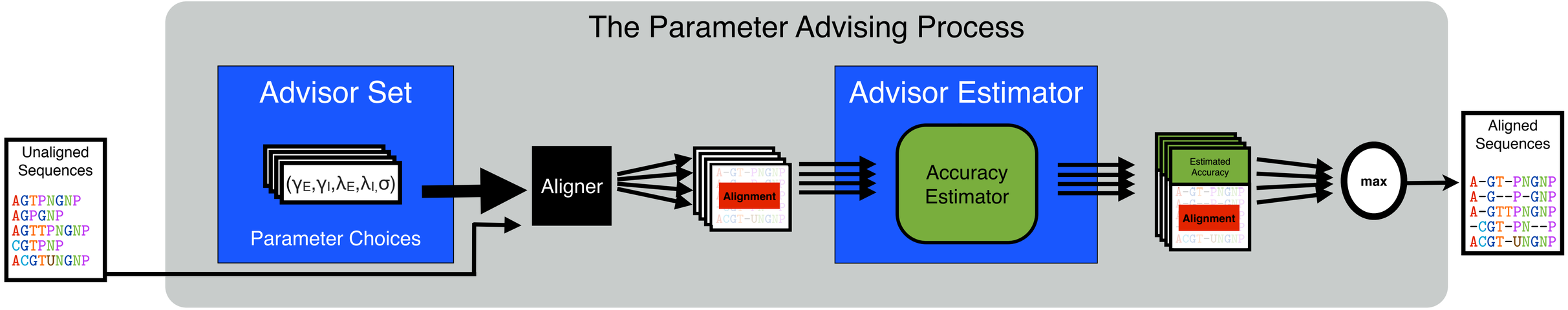
- an advisor set, that is
- optimal for an oracle advisor,
- can be found by integer linear programming,
- can work well with an actual advisor.

[ACM-BCB 2014]

Parameter advising

A parameter advisor has two components:

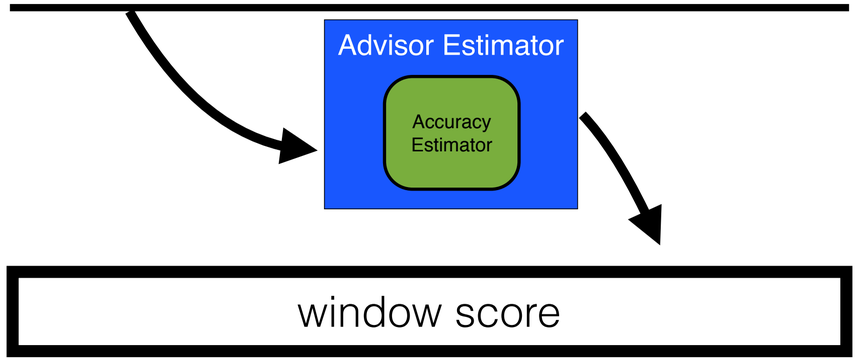
- an accuracy estimator, and
- a set of candidate parameter choices.



[RECOMB 2012]

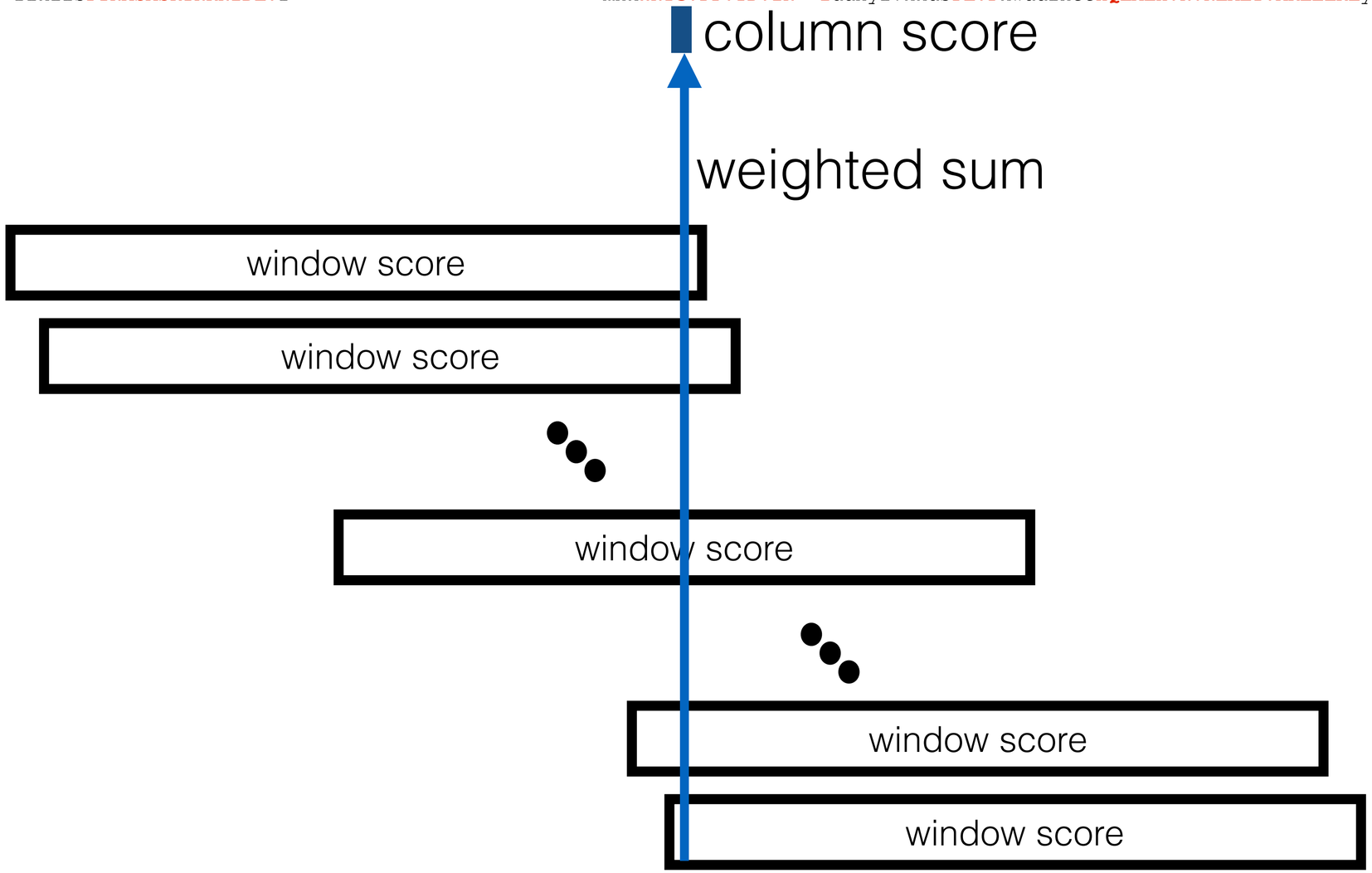
Adaptive local realignment

```
rkeyagLYHEVAQAHGVDVSQVrqMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnndPLVFrwddsnaqHKLTLLVNQNVDGEAARAEARVyleefvresysnt
kkaqldLYNEVATEHGYDVTKId-MKFGNFLLFDTVWLlehhftefgllldqmskgrfrfydlmkegfnegyiaadnePMILswiintheHCLSYITSVDHDSNRAKDICRNflghwy-dsyvna
rielln HYQAAAAKFNVDIANVr--------------------------mtkWNYGVFFLYDVVA--Fsehhidksyn PLLFkwddsqqkHRLMLFVNVNDNPTQAKAELSIyledyl--sytqa
rlkllsFYNASASKYNKNIDLVr--------------------------mnkWNYGVFFVYDVIN--Iddhylvkkds PLVFkwddineeHQLMLHVNVNEAETVAKEELKLyienyv--actqp
```



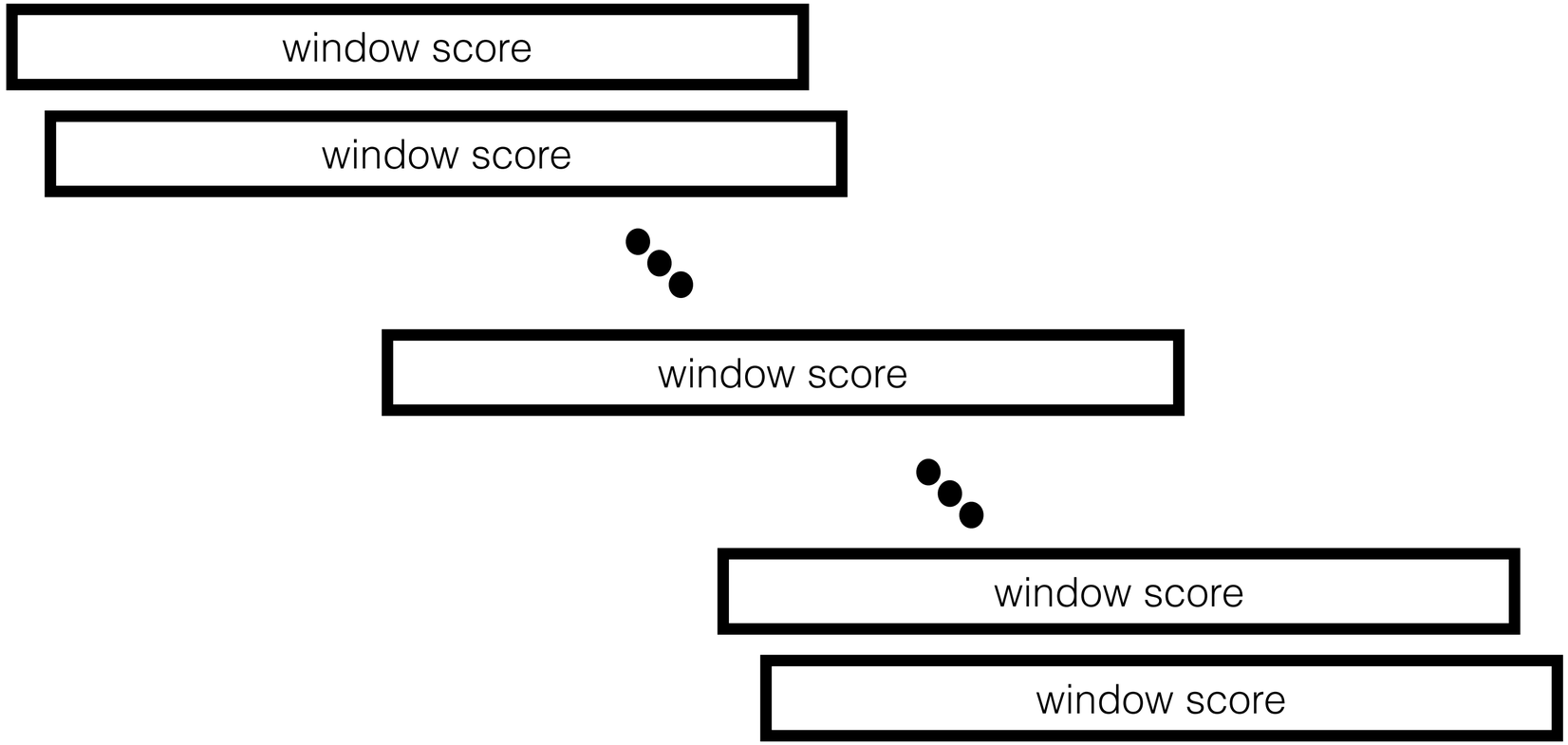
Adaptive local realignment

```
rkeyagLYHEVAQAHGVDVSQVrqMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnndPLVFrwddsnaqHKLTLLVNQNVDGEAARAEARVyleefvresysnt
kkaqldLYNEVATEHGYDVTKId-MKFGNFLLFDTVWLlehhftefgllldqmskgrfrfydlmkegfnegyiaadnePMILswiintheHCLSYITSVDHDSNRAKDICRNflghwy-dsyvna
rielln HYQAAAAKFNVDIANVr--------------------------mtkWNYGVFFLYDVVA--Fsehhidksyn PLLFkwddsqqkHRLMLFVNVNDNPTQAKAELSIyledyl--sytqa
rlkllsFYNASASKYNKNIDLVr--------------------------mnkWNYGVFFVYDVIN--Iddhylvkkds PLVFkwddineeHQLMLHVNVNEAETVAKEELKLyienyv--actqp
```

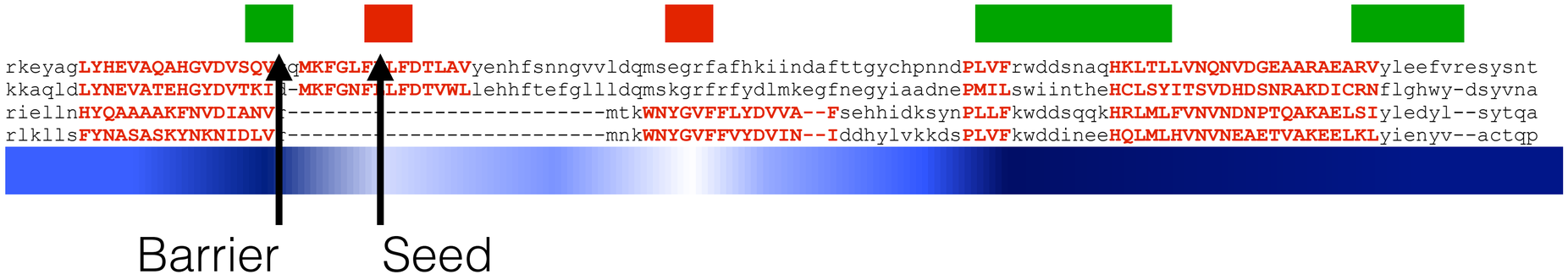


Adaptive local realignment

```
rkeyagLYHEVAQAHGVDVSQVrqMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnndPLVFrwddsnaqHKLTLLVNQNVDGEAARAEARVyleefvresysnt
kkaqldLYNEVATEHGYDVTKId-MKFGNFLLFDTVWLlehhftefgllldqmskgrfrfydlmkegfnegyiaadnePMILswiintheHCLSYITSVDHDSNRAKDICRNflghwy-dsyvna
rielln HYQAAAAKFNVDIANVr--------------------------mtkWNYGVFFLYDVVA--Fsehhidksyn PLLFkwddsqqkHRLMLFVNVNDNPTQAKAELSIyledyl--sytqa
rlkllsFYNASASKYNKNIDLVr---------------------------mnkWNYGVFFVYDVIN--Iddhylvkkds PLVFkwddineeHQLMLHVNVNEAETVAKEELKLyienyv--actqp
```



Adaptive local realignment

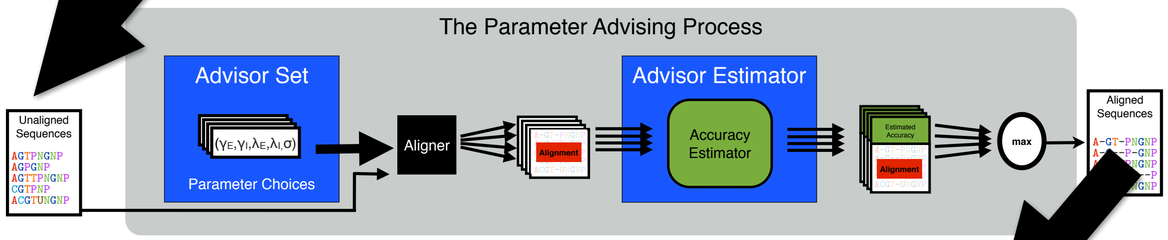


Adaptive local realignment

Adaptive local realignment

```
rkeyagLYHEVAQAHGVDVSQVrqMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnndRLVFrwddsnaqHKLTLLVNQNVDGEAARAEARVyleefvresysnt
kkaqldLYNEVATEHGYDVTKId-MKFGNFLLFDTVWLlehhftefgllldqmskgrfrfydlmkegfnegyiaadneRMILswiintheHCLSYITSVDHDSNRAKDICRNflghwy-dsyvna
rielln HYQAAAAKFNVDIANVr-------------------------mtkWNYGVFFLYDVVA--Fsehhidksyn RLLFkwddsqqkHRLMLFVNVNDNPTQAKAELSIyledyl--sytqa
rlkllsFYNASASKYNKNIDLVr-------------------------mnkWNYGVFFVYDVIN--Iddhylvkkds RLVFkwddineeHQLMLHVNVNEAETVAKEELKLyienyv--actqp
```

```
qMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnnd
-MKFGNFLLFDTVWLlehhftefgllldqmskgrfrfydlmkegfnegyiaadne
--------------------------mtkWNYGVFFLYDVVA--Fsehhidksyn
--------------------------mnkWNYGVFFVYDVIN--Iddhylvkkds
```




```
--qMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnnd
---MKFGNFLLFDTVWLlehhftefgllldqmskgrfrfydlmkegfnegyiaadne
mtkWNYGVFFLYDVVAFsehhidksyn------------------------------
mnkWNYGVFFVYDVINIddhylvkkds------------------------------
```

```
rkeyagLYHEVAQAHGVDVSQVr--qMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnnd RLVFrwddsnaqHKLTLLVNQNVDGEAARAEARVyleefvresysnt
kkaqldLYNEVATEHGYDVTKIc---MKFGNFLLFDTVWLlehhftefgllldqmskgrfrfydlmkegfnegyiaadne RMILswiintheHCLSYITSVDHDSNRAKDICRNflghwy-dsyvna
rielln HYQAAAAKFNVDIANVrmtkWNYGVFFLYDVVAFsehhidksyn------------------------------ RLLFkwddsqqkHRLMLFVNVNDNPTQAKAELSIyledyl--sytqa
rlkllsFYNASASKYNKNIDLVrmnkWNYGVFFVYDVINIddhylvkkds------------------------------ RLVFkwddineeHQLMLHVNVNEAETVAKEELKLyienyv--actqp
```

Adaptive local realignment

Accuracy
33%

```
qMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnnd
-MKFGNFLLFDTVWLlehhftefglllldqmskgrfrfydlmkegfnegyiaadne
--------------------------mtkWNYGVFFLYDVVA--Fsehhidksyn
--------------------------mnkWNYGVFFVYDVIN--Iddhylvkkds
```

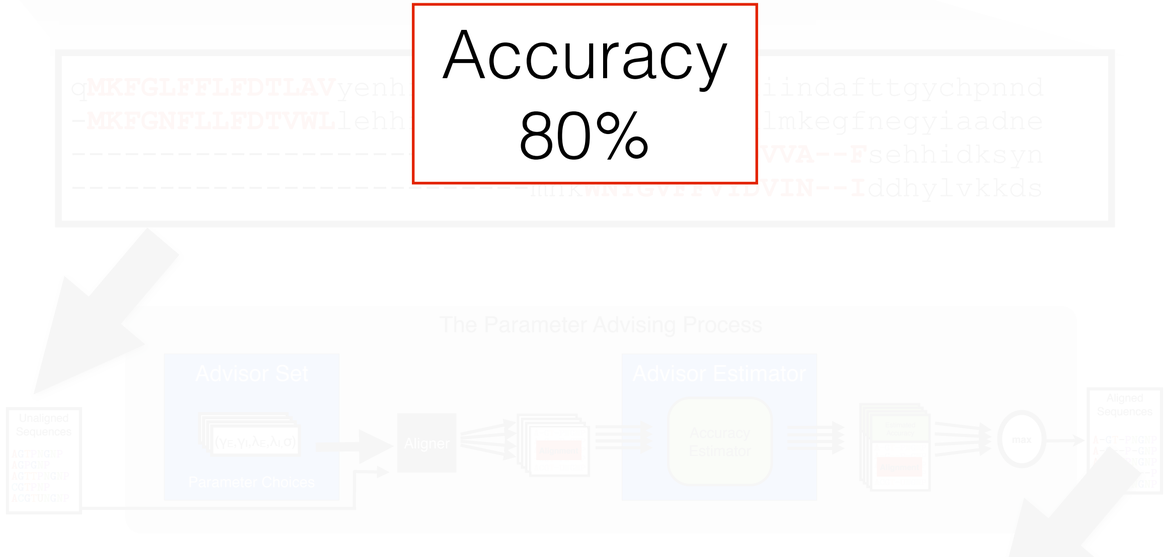
```
--qMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnnd
---MKFGNFLLFDTVWLlehhftefglllldqmskgrfrfydlmkegfnegyiaadne
mtkWNYGVFFLYDVVAFsehhidksyn------------------------------
mnkWNYGVFFVYDVINIddhylvkkds------------------------------
```

Accuracy
100%

Adaptive local realignment

```
rkeyagLYHEVAQAHGVDVSQVrqMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnndRLVFrwddsnaqHKLTLLVNQNVDGEAARAEARVyleefvresysnt
kkaqldLYNEVATEHGYDVTKId-MKFGNFLLFDTVWLlehhftefgllldqmskgrfrfydlmkegfnegyiaadneRMILswiintheHCLSYITSVDHDSNRAKDICRNflghwy-dsyvna
rielln HYQAAAAKFNVDIANVr-------------------------mtkWNYGVFFLYDVVA--Fsehhidksyn RLLFkwddsqqkHRLMLFVNVNDNPTQAKAELSIyledyl--sytqa
rlkllsFYNASASKYNKNIDLVr-------------------------mnkWNYGVFFVYDVIN--Iddhylvkkds RLVFkwddineeHQLMLHVNVNEAETVAKEELKLyienyv--actqp
```

Accuracy
80%



Accuracy
100%

```
rkeyagLYHEVAQAHGVDVSQVr--qMKFGLFFLFDTLAVyenhfsnngvvldqmsegrfafhkiindafttgychpnndRLVFrwddsnaqHKLTLLVNQNVDGEAARAEARVyleefvresysnt
kkaqldLYNEVATEHGYDVTKIc---MKFGNFLLFDTVWLlehhftefgllldqmskgrfrfydlmkegfnegyiaadneRMILswiintheHCLSYITSVDHDSNRAKDICRNflghwy-dsyvna
rielln HYQAAAAKFNVDIANVrmtkWNYGVFFLYDVVAFsehhidksyn-----------------------------RLLFkwddsqqkHRLMLFVNVNDNPTQAKAELSIyledyl--sytqa
rlkllsFYNASASKYNKNIDLVrmnkWNYGVFFVYDVINIddhylvkkds-----------------------------RLVFkwddineeHQLMLHVNVNEAETVAKEELKLyienyv--actqp
```

Adaptive local realignment

Adaptive local realignment takes as input

- an initial alignment,
- an advisor set,
- window sizes, and
- column score thresholds.

And outputs

- a new alignment
- constructed by realigning regions of low estimated accuracy.

Experimental results

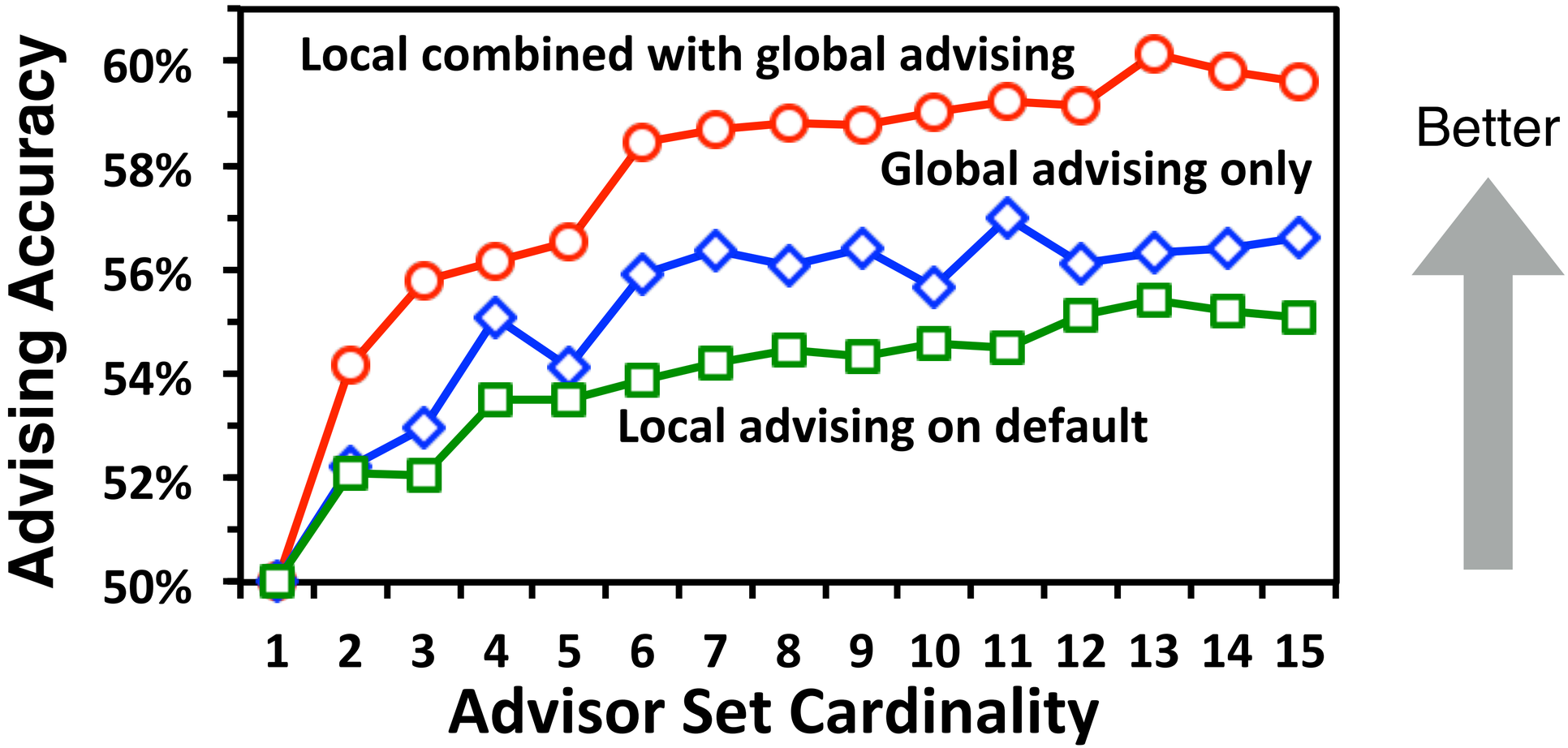
We correct for the bias in over-representation of easy-to-align benchmarks.

- The difficulty of a benchmark is its accuracy under the default parameter setting.
- Split the range of difficulties [0,1] into 10 bins.
- Report advisor accuracy uniformly averaged across bins.

The typical average accuracy is close to 50%.

Experimental results

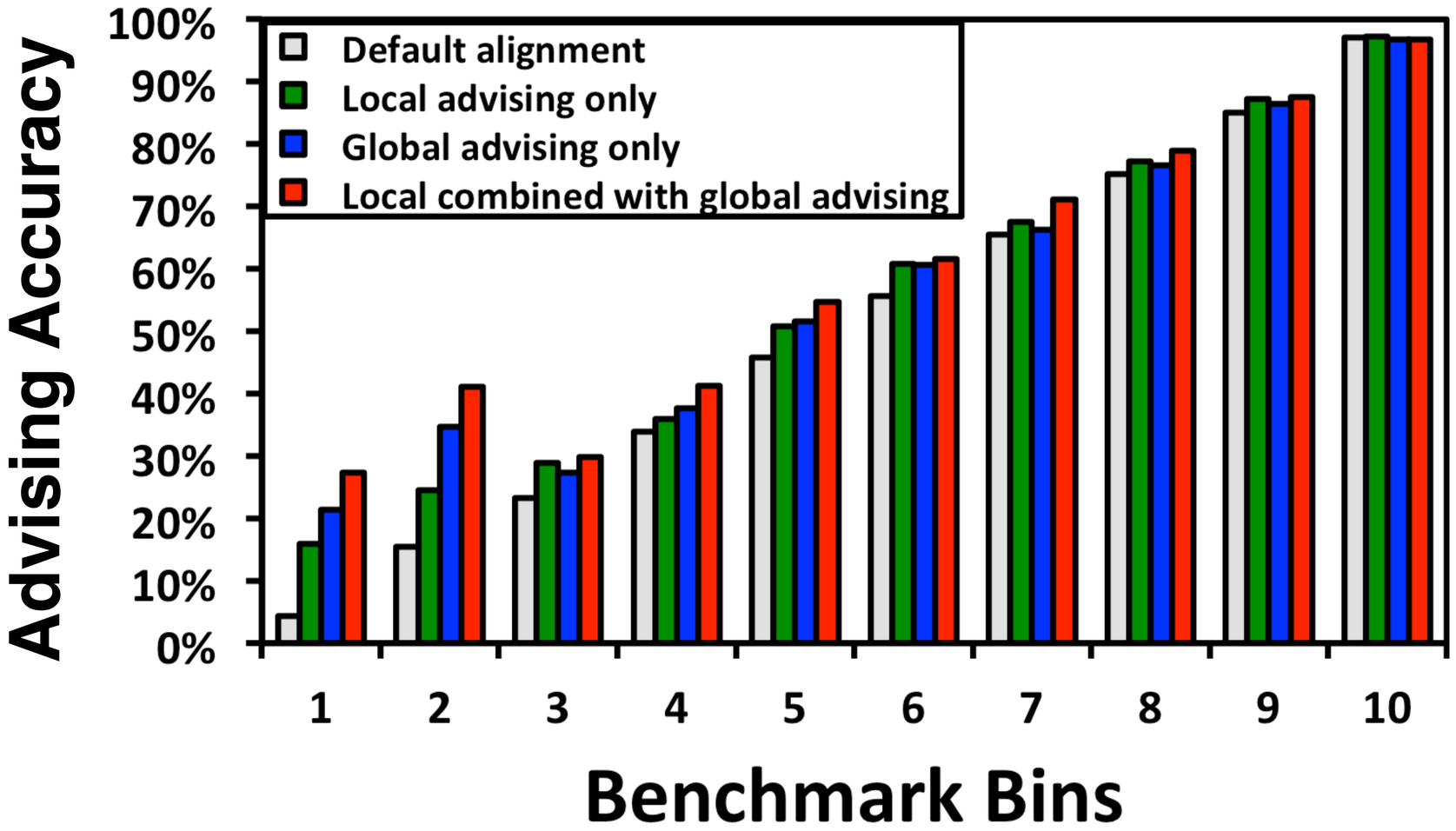
Average accuracy of advisors versus set cardinality



Boosts accuracy by 9% over default parameter choice

Experimental results

Average accuracy of advisors within difficulty bins



Better

oracle advising set
$k = 10$

Boosts accuracy on the hardest bins by almost 26%

Conclusions

Local adaptive realignment:

- Uses diverse parameter choices for heterogeneous proteins.
- `Facet` estimator identifies misaligned regions.
- Misaligned regions are polished with parameter advising.
- With global advising, boosts accuracy by up to 26%.

Further research

Promising directions include:

- Finding advisor sets tuned for local realignment.
- Improving the estimator by training on local examples.
- Using an ensemble of aligners to realign regions.

Software distribution

Available for download:

- `Facet` accuracy estimator
- `Opal` aligner with global and local parameter advising
- Parameter sets for advising

`facet.cs.arizona.edu`

Acknowledgments

People

William Pearson

Travis Wheeler

Carl Kingsford

Funding

University of Arizona

- NSF IGERT in Genomics Grant DGE-0654435
- NSF Grant IIS-1217886

Carnegie Mellon University

- NSF Grants CCF-1256087 & CCF-131999
- NIH Grant R01HG007104
- Gordon and Betty Moore Foundation Grant GBMF4554

The University of Arizona | Computer Science UA Science

Computational Biology Department | Carnegie Mellon University School of Computer Science

NSF

National Institutes of Health

Gordon and Betty Moore Foundation