Ensemble Multiple Sequence Alignment via Advising

Dan DeBlasio

John Kececioglu

Department of Computer Science
University of Arizona

Motivation

Multiple sequence alignment is a fundamental problem in bioinformatics.

- standard multiple sequence alignment is NP-Complete
- many popular aligners for multiple sequence alignment
- each aligner has many parameters whose values affect the accuracy of the alignment

Opal

```
... ------mlpgtffevlkne-----gvvAIATQg-edgph--lvntwnsylk---vldg ...
... d-dpidlftkwfneakedpretlpeaiTFSSAelpsgr----vssrillfk---eldh ...
... sldpvkqfaawfeeavqcpdigeanamCLATCt-rdgk----psarmlllk---gfgk ...
... s-mdfedfpvesahriltpr----ptvMVTTVd-eegn----inaapfsftmpvsidp ...
... --mdveafykisy-----------glyIVTSE-sngrkcgqiantvfqlt---s-kp ...
```

MUMMALS

```
... fevlknegvvAIATQgedgphlvntwnsylkv-ldgnrivvpvggmhkteanva-rde ...
... tkw-fn--------eakedpret-------------lpeaiTFSS-------Aelpsg ...
... aaw-fe--------eavqcpdig-------------eanamCLAT-------Ct-rdg ...
... hriltprptvMVTTVdeegninaapfsftmpvsidppvvafasapdhhtarnie-sth ...
... ykisyglyIVTSEsngrkcgqiant---vfqltskpvqiavclnkendthnavk-esg ...
```

Motivation

How do we combine a collection of aligners and parameter choices into a new alignment method that is better than any single choice?

We approach this question through the framework of advising.

Advising

Advising for input sequences S is

- selecting the aligner $\mathbb{A}$ and parameter choice p from a set of pairs P
- that produces the alignment with highest estimated accuracy E.

$$\text{Advice}_P(S) \quad := \quad \underset{(\mathbb{A},p)\ \in\ P}{\text{argmax}}\ E\Big(\mathbb{A}_p(S)\Big)$$

Advising variants

- General aligner advising [BCB'15]
  - Selecting from a set of aligners and multiple parameter settings.
- Default aligner advising [BCB'15]
  - Selecting from a set of aligners that use their default parameter setting.
- Parameter advising [Kececioglu and DeBlasio 2013]
  - Selecting from a set of parameter choices for a single aligner.

Advising variants

- General aligner advising [BCB'15]
  - Selecting from a set of aligners and multiple parameter settings.
- Default aligner advising [BCB'15]
  - Selecting from a set of aligners that use their default parameter setting.

Default advising and general advising yield two forms of ensemble alignment.

Advising

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 N H l t …	… a N – H l t …
↑ ↑	

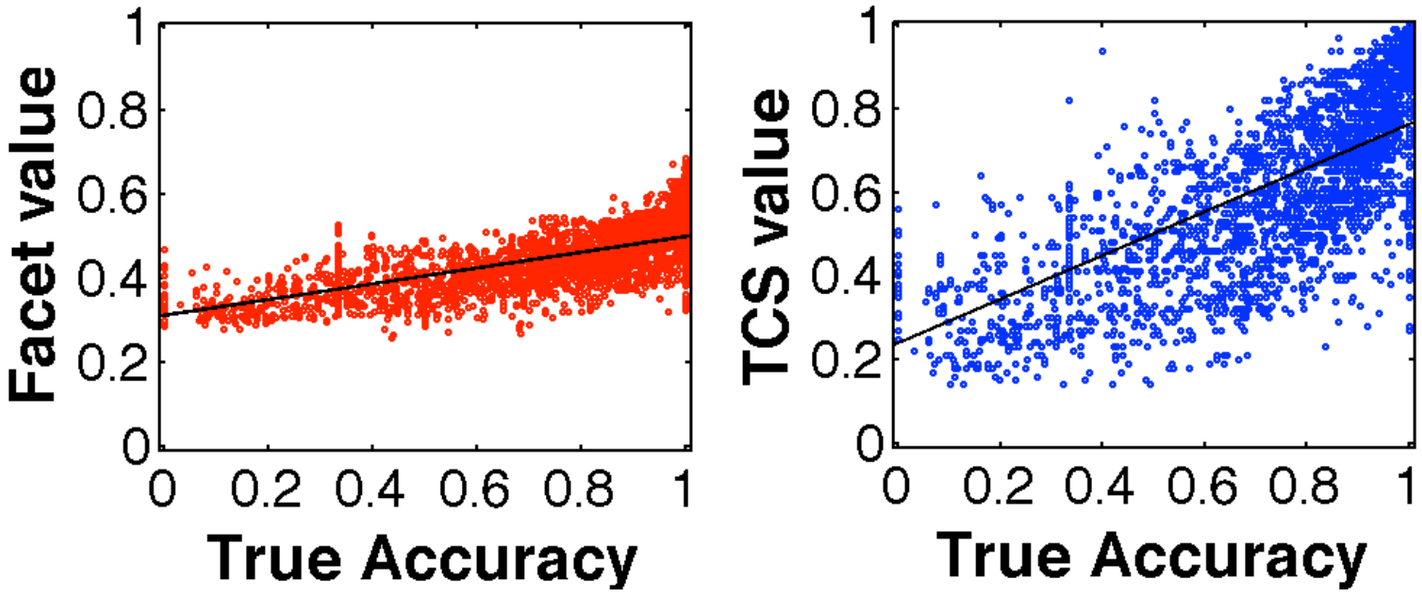
66% Accuracy

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

Accuracy estimators

The best estimators of alignment accuracy *without a reference* include:

- `MOS` [Lassmann and Sonnhammer, 2005]
- `PredSP` [Ahola, *et al.*, 2008]
- `Guidance` [Penn, *et al.*, 2010]
- `Facet` [Kececioglu and DeBlasio, 2013]
- `TCS` [Chang, Tommaso and Notredame, 2014]



Accuracy estimators

The best estimators of alignment accuracy *without a reference* include:

- `MOS` [Lassmann and Sonnhammer, 2005]
- `PredSP` [Ahola, *et al.*, 2008]
- `Guidance` [Penn, *et al.*, 2010]
- `Facet` [Kececioglu and DeBlasio, 2013]
- `TCS` [Chang, Tommaso and Notredame, 2014]

An oracle is a *perfect* advisor whose "estimator" is true accuracy.

Advising

An advisor has two components:

- an accuracy estimator $E(A)$, and
- a set of aligner and parameter choice pairs P.

Given accuracy estimator E,
what is the optimal set P
of pairs?

Advisor Set problem

For the Advisor Set problem the input is

- universe of aligners and parameters choices U,
- cardinality bound k,
- estimator values, accuracies, and weights for all examples.

Advisor Set problem

The output is

- an optimal set $P \subseteq U$ of aligners and parameter choices with $|P| \leq k$, that maximizes the average advising accuracy

$$\sum_{\text{Benchmark } i} w_i \ \text{Accuracy}\Big(\text{Advice}_P\,(S_i) \ \Big)$$

Advisor Set problem

THEOREM [DeBlasio and Kececioglu 2014]

The Advisor Set problem is NP-complete.

- Polynomial-time solvable for fixed k
- Optimal oracle sets can be found in practice for very large k

Approximation algorithm

THEOREM [DeBlasio and Kececioglu 2014]

There is an efficient greedy $\frac{\ell}{k}$-approximation algorithm for Advisor Set, for any fixed $\ell \leq k$.

Related work

- `AQUA` [Muller, Creevey, Thompson, Arendt, and P. Bork 2010]
  - Chooses between `MAFFT` and `MUSCLE` alignments of the same sequences using `NorMD` values.

- `M-Coffee` [Wallace, O'Sullivan, Higgins, and Notredame 2006]
  - Aligns sequences using `T-Coffee`, whose scoring function combines the outputs of several aligners.
  - The authors call this approach meta-alignment.

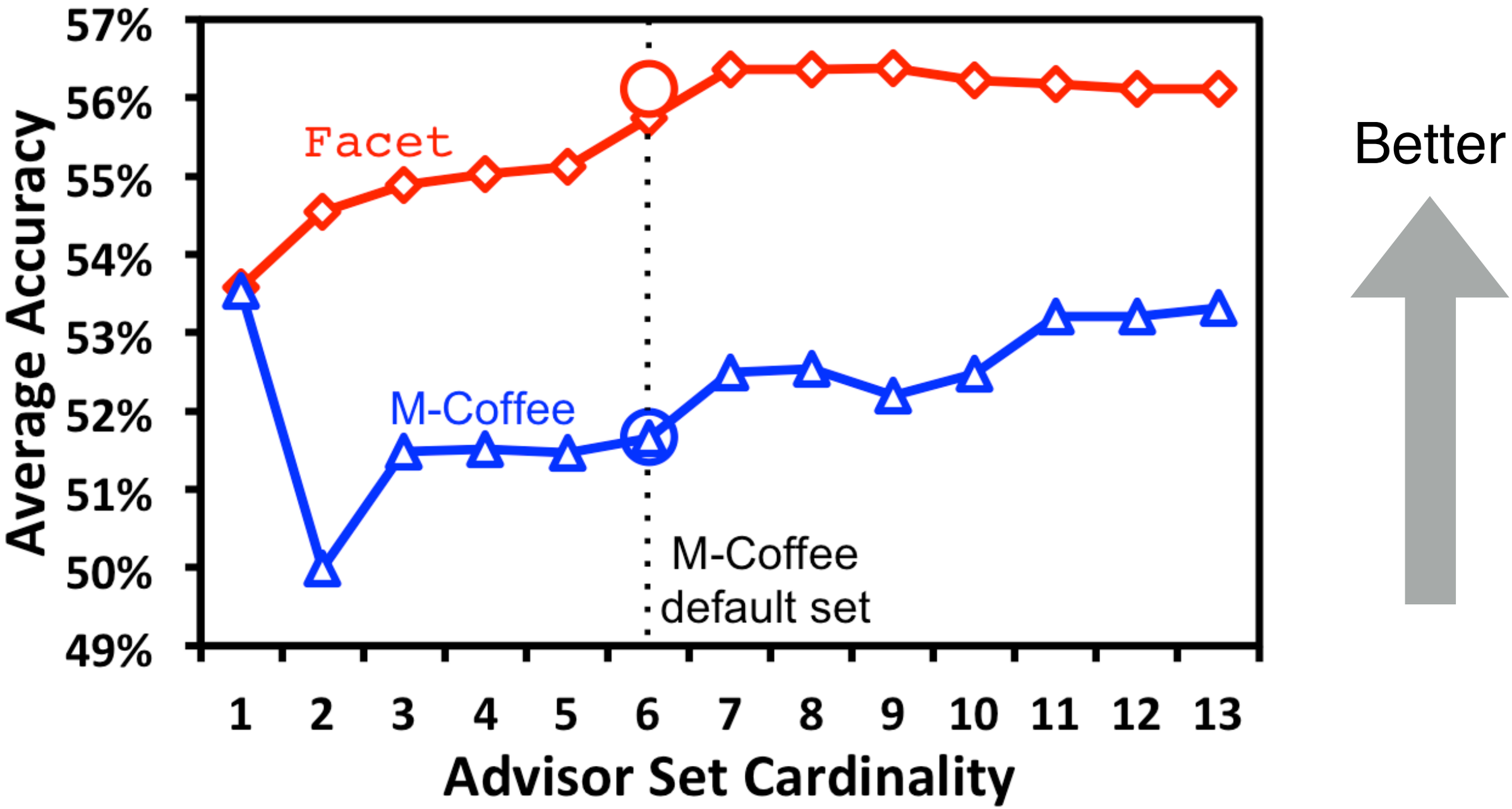
Experimental results

We compare ensemble alignment using `Facet` to meta-alignment using `M-Coffee`.

- Used the 13 aligners included in `M-Coffee`.
- Found oracle sets of these aligners.
- Compared `Facet` and `M-Coffee` on the same oracle sets.
- The default setting for `M-Coffee` has 6 aligners.

Experimental results

Average accuracy versus set cardinality



Ensemble alignment significantly improves on meta-alignment

Default aligner advising

The universe for default aligner advising includes

- the most commonly-used aligners (17 tools),
- using their default parameter settings.

Default aligner advising

The 17 aligners are:

- `ClustalW` (1994)
- `ClustalW2` (2007)
- `Clustal Omega` (2011)
- `DIALIGN-TX` (2008)
- `FSA` (2009)
- `Kalign` (2005)
- `MAFFT` (2005)
- `MUMMALS` (2006)
- `MUSCLE` (2004)
- `MSAProbs` (2010)
- `Opal` (2007)
- `POA` (2002)
- `PRANK` (2005)
- `Probalign` (2006)
- `ProbCons` (2005)
- `SATé` (2011)
- `T-Coffee` (2000)

Default aligner advising

The 17 aligners are:

- ClustalW (1994)
- ClustalW2 (2007)
- Clustal Omega (2011)
- DIALIGN-TX (2008)
- FSA (2009)
- Kalign (2005)
- MAFFT (2005)
- MUMMALS (2006)
- MUSCLE (2004)
- MSAProbs (2010)
- Opal (2007)
- POA (2002)
- PRANK (2005)
- Probalign (2006)
- ProbCons (2005)
- SATé (2011)
- T-Coffee (2000)

General aligner advising

Constructed a parameter universe for 10 aligners

- by finding their tunable parameters,

for the `Opal` aligner, a parameter choice is a 5-tuple

$$(\sigma, \gamma_I, \gamma_E, \lambda_I, \lambda_E)$$

General aligner advising

Constructed a parameter universe for 10 aligners

- by finding their tunable parameters,
- for numerical parameters, values cover range,
- for discrete parameters, enumerate all choices,
- goal of 100 parameter settings for each aligner.

We combine each of these with default aligner advising universe for general aligner advising.

Experimental results

We evaluate the accuracy of advising

- with the `Facet` and `TCS` estimators,
- consider greedy advisor sets for both universes,
- on over 800 benchmarks from `BENCH` and `PALI`,
- using 12-fold cross-validation.

Experimental results

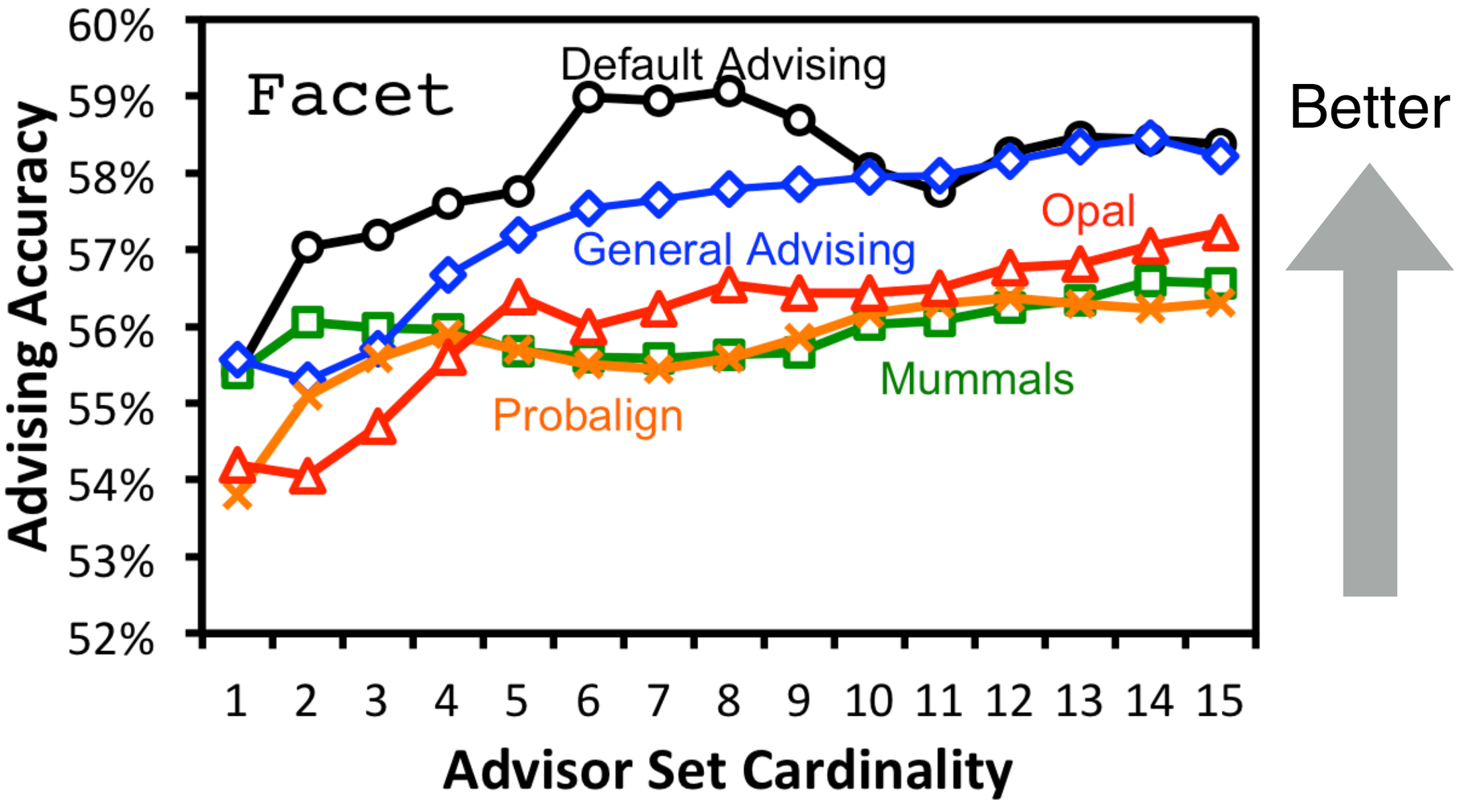
We correct for the bias in over-representation of easy-to-align benchmarks.

- The difficulty of a benchmark is its average accuracy under the default parameter setting for `Clustal Omega`, `MAFFT`, and `ProbCons`.
- 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

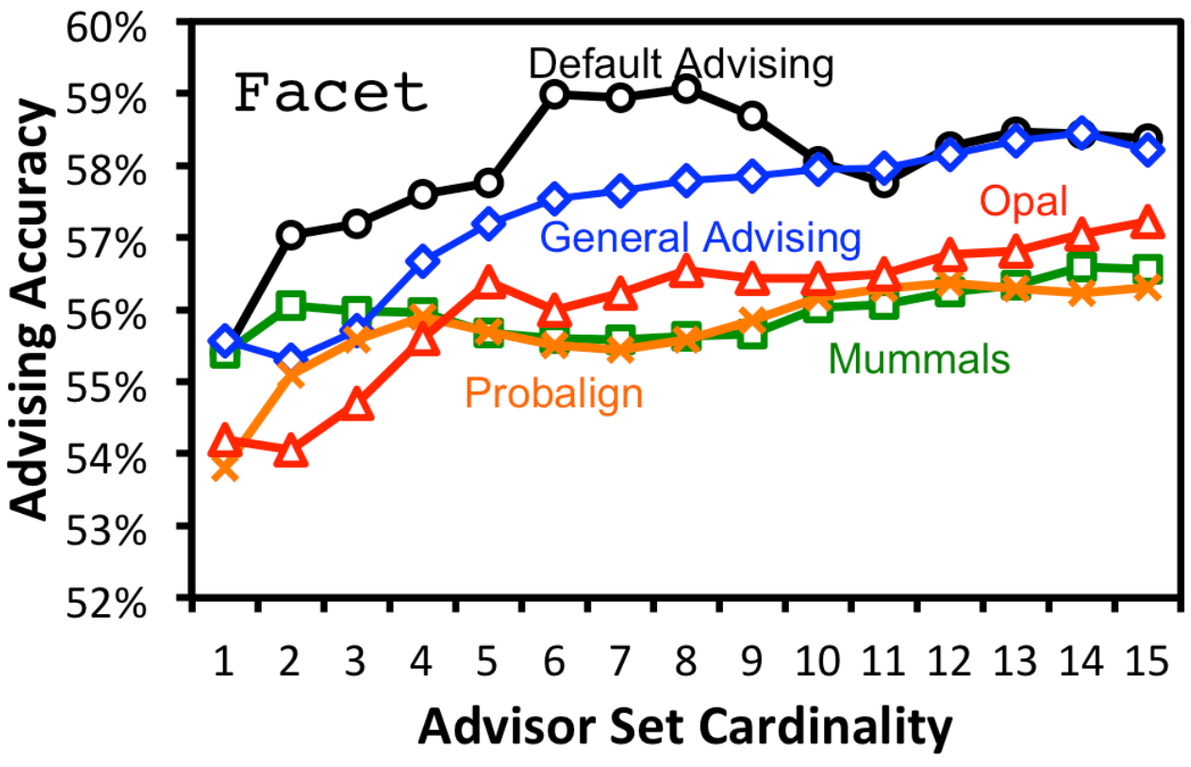
Advisor performance versus set cardinality

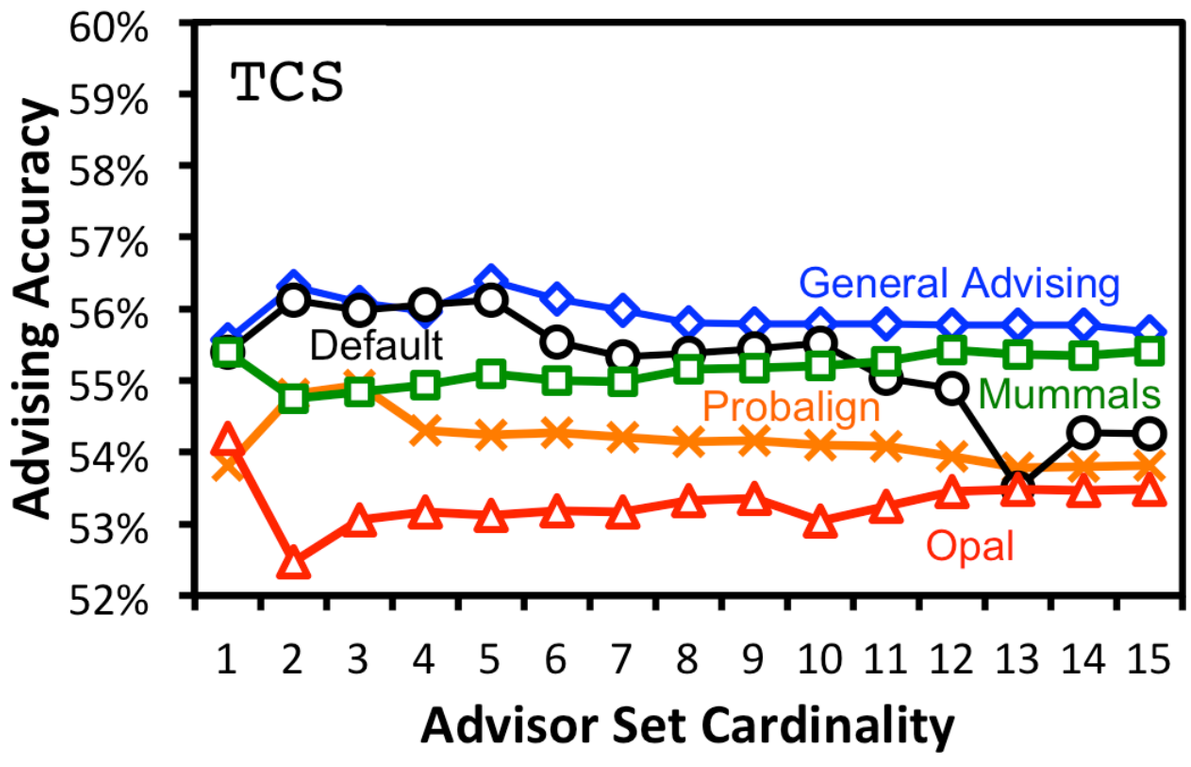


Ensemble advising boosts accuracy over parameter advising

Experimental results

Advisor performance versus set cardinality

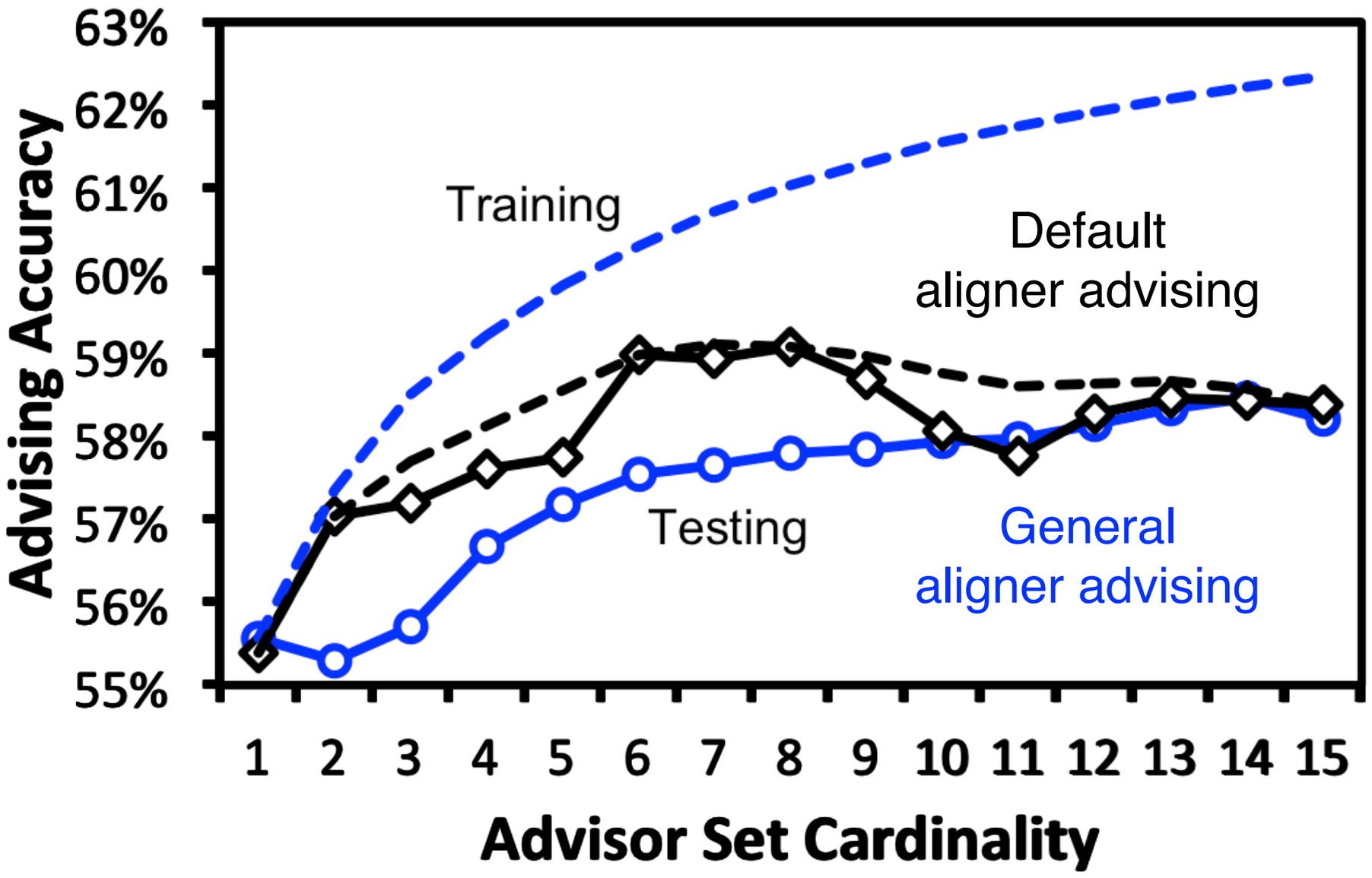




Facet outperforms `TCS` accuracy estimator

Experimental results

Advisor performance versus set cardinality



Default aligner advising sets generalize better

Conclusions

Ensemble alignment significantly increases accuracy.

- Advising yields the first successful ensemble method for alignment.
- Parameter advising boosts accuracy for nearly all standard aligners.
- Aligner advising further improves upon parameter advising.

Further research

Future directions for ensemble alignment include:

- Learning advisor sets with improved generalization
- Developing more accurate estimators
- Extending to aligning DNA and RNA sequences

Software distribution

Available for download:

- `Facet` estimator
- Ensemble alignment tool
- Precomputed ensemble sets for all aligners
- Benchmark suites with structure predictions

`facet.cs.arizona.edu`

Acknowledgments

People

William Pearson

Travis Wheeler

Funding

- University of Arizona
  NSF IGERT in Genomics
  Grant DGE-0654435
- NSF Grant IIS-1217886