Learning Advisors for Multiple Sequence Alignment

Dan DeBlasio
John Kececioglu

Department of Computer Science, University of Arizona

ARIZONA

COMPUTER SCIENCE
UA SCIENCE

ISCB Student COUNCIL
10 YEARS
SYMPOSIUM 2014

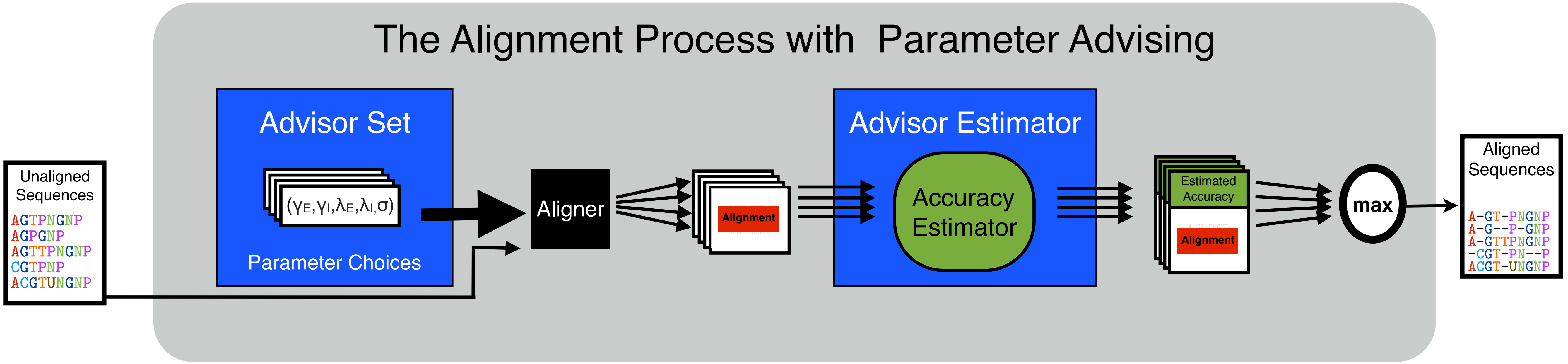
Parameter advising

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

- The default attempts to have good *average* accuracy across benchmarks.
- An optimal default choice can be found by inverse alignment [Kececioglu and Kim 2007].
- The default may be a poor choice for specific inputs.

Can we boost aligner accuracy
by an input-dependent choice
of parameter values?

Parameter advising

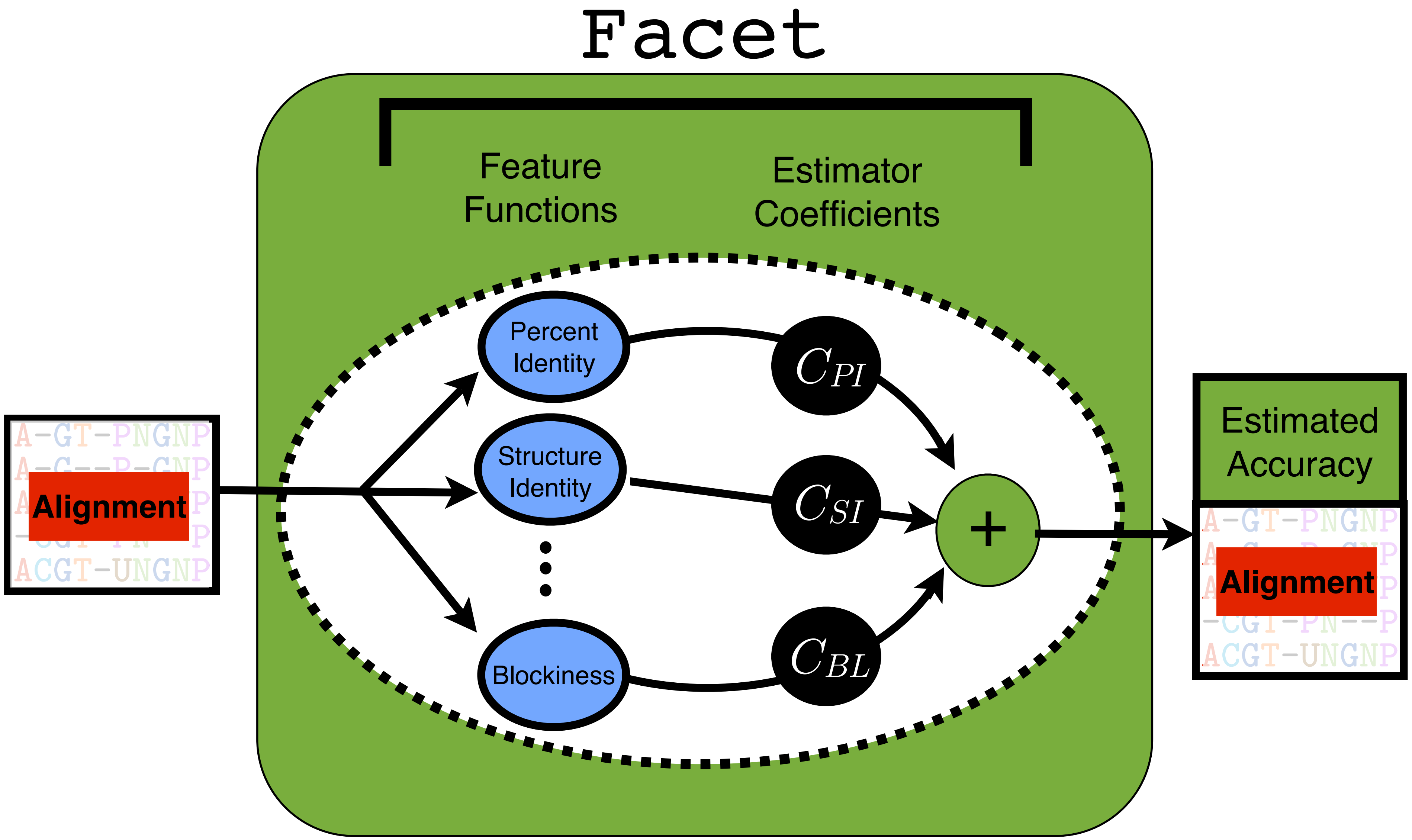


An advisor has two ingredients:

(1) the advisor set of parameter choices used to generate candidate alignments, and

(2) an advisor estimator that ranks alignments by estimated accuracy.

Accuracy estimator



Our accuracy estimator Facet (**F**eature-based **Ac**curacy **Est**imator) is

- a linear combination
- of real–valued feature functions

Parameter advising

Parameter advising is selecting a parameter choice p from a set P to maximize the accuracy of an aligner $\mathcal{T}$.

- Given estimator E_c , an advisor finds a parameter choice $\tilde{p}$ for input sequences S.

$$\tilde{p} \quad := \quad \underset{p \,\in\, P}{\operatorname{argmax}} \; E_c\Big(\mathcal{T}_p(S)\Big)$$

- The oracle is a perfect advisor that uses true accuracy.

Problems

Finding a parameter advisor involves solving two problems:

- learning advisor coefficients, and
- finding a advisor set of parameter choices.

Problems

There is an issue with defining the accuracy of an advisor when there are ties in estimator value:

- In practice the advisor selects among the alignments that have maximum estimator value.
- When learning an advisor we want to maximize the expected accuracy.

Problems

We learn the estimator using examples consisting of

- an alignment A_{ij} produced by aligning benchmark i using parameter choice j,
- the associated feature vector F_{ij} = $F(A_{ij})$,
- the true accuracy a_{ij} of A_{ij}.

To correct for bias in easy benchmarks we assign a weight w_i to each.

Problems

A parameter choice i consists of an assignment of the values of the alignment parameters.

- For Opal a parameter choice is a 5-tuple

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

- The universe U is a collection of these parameter choices.

Problems

- The potential output set of parameter choices for the advisor on benchmark i with parameter set P is

$$\mathcal{O}_i(P) \;:=\; \Big\{ j \in P \;:\; E_c(A_{ij}) \;\geq\; e_i^* - \epsilon \Big\}$$

where

$$e_i^* := \max \Big\{ E_c(A_{i\tilde{j}}) : \tilde{j} \in P \Big\}$$

- The expected accuracy of the advisor is the average accuracy over these parameter choices

$$\mathcal{A}_i(P) := \frac{1}{|\mathcal{O}_i(P)|} \sum_{j \in \mathcal{O}_i(P)} a_{ij}$$

Advisor Sets

The input to the Advisor Set problem is

- cardinality bound k,
- benchmark weights w_i, where $\sum_i w_i = 1$, 0 ≤ w_i ≤ 1
- accuracies a_{ij}, where 0 ≤ a_{ij} ≤ 1
- feature vectors F_{ij} = $(f_{ij1}, f_{ij2}, \cdots, f_{ijt})$, where 0 ≤ f_{ijh} ≤ 1
- error tolerance ε≥0
- estimator coefficients c = (c_1 , ... , c_t), where each c_i ≥ 0 and $\sum_i c_i = 1$, and
- universe of parameters choices U.

Advisor Sets

The output is

- a set $P \subseteq U$ of parameter choices, where $|P| \leq k$ that maximizes the objective function

$$\sum_i w_i \mathcal{A}_i(P)$$

The Advisor Set problem is NP-complete.

Finding advisor sets

Advisor Set can be modeled as an integer linear program.

- ILP cannot be solved to optimality in a reasonable amount of time.
- Optimal sets for small cardinalities k can be found by exhaustive search.

We have an approximation algorithm that

- finds an $\frac{l}{k}$-approximation of the optimal advisor set,
- for any constant $l \leq k$.

The approximation ratio is tight for tolerance ε = 0.

Advisor Estimator

The input to the Advisor Estimator problem is

- weights w_i on the benchmarks,
- accuracies a_{ij} of the alternate alignments,
- feature vectors F_{ij} for the alternate alignments,
- error tolerance ε, and
- advisor set P of parameter choices.

Advisor Estimator

The output is

- estimator coefficient vector c = (c_1 , ... , c_t), where each c_i ≥ 0 and $\sum_i c_i = 1$ that maximizes the objective function

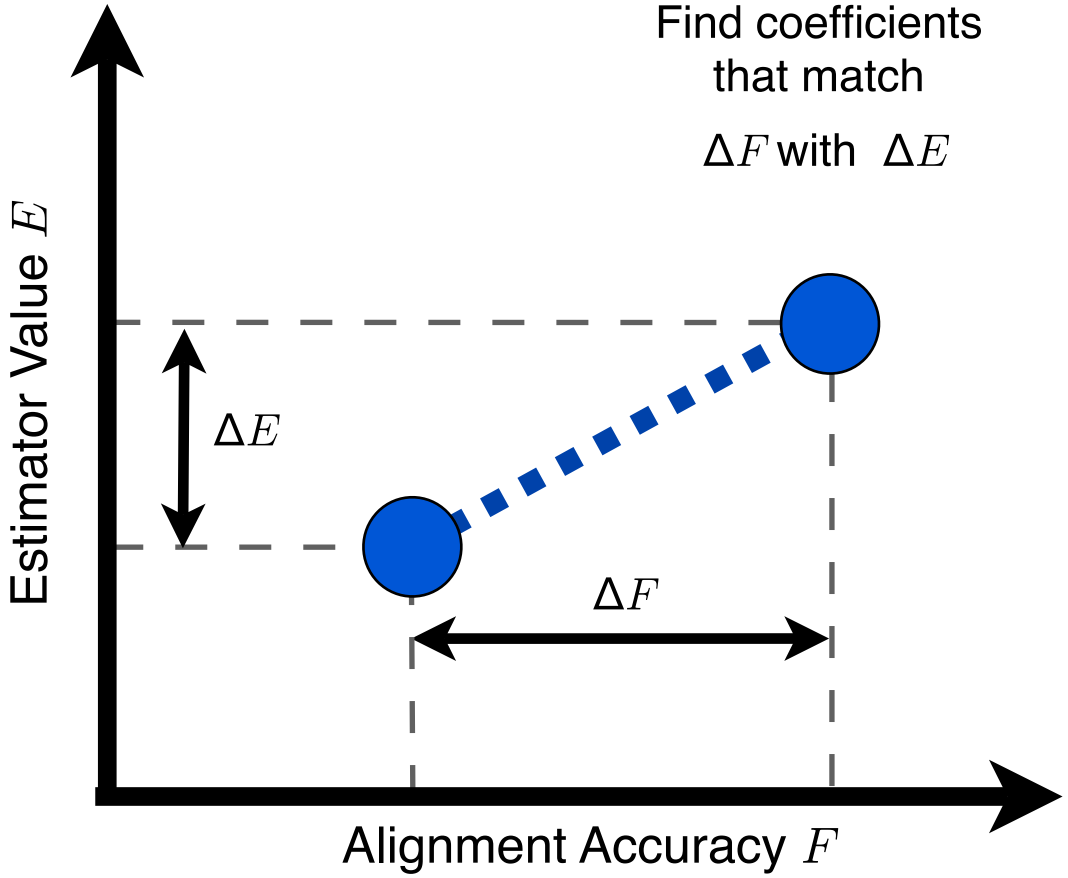
$$\sum_i w_i \mathcal{A}_i(P)$$

The Advisor Estimator problem is NP-complete.

Learning the estimator

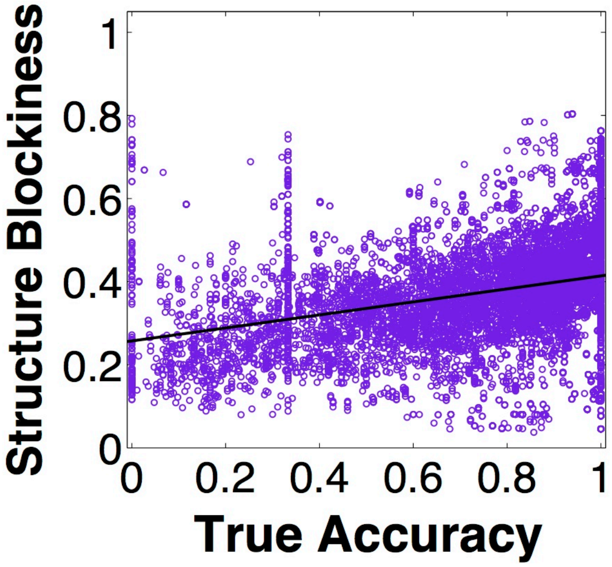
To learn the estimator we find optimal coefficients that fit

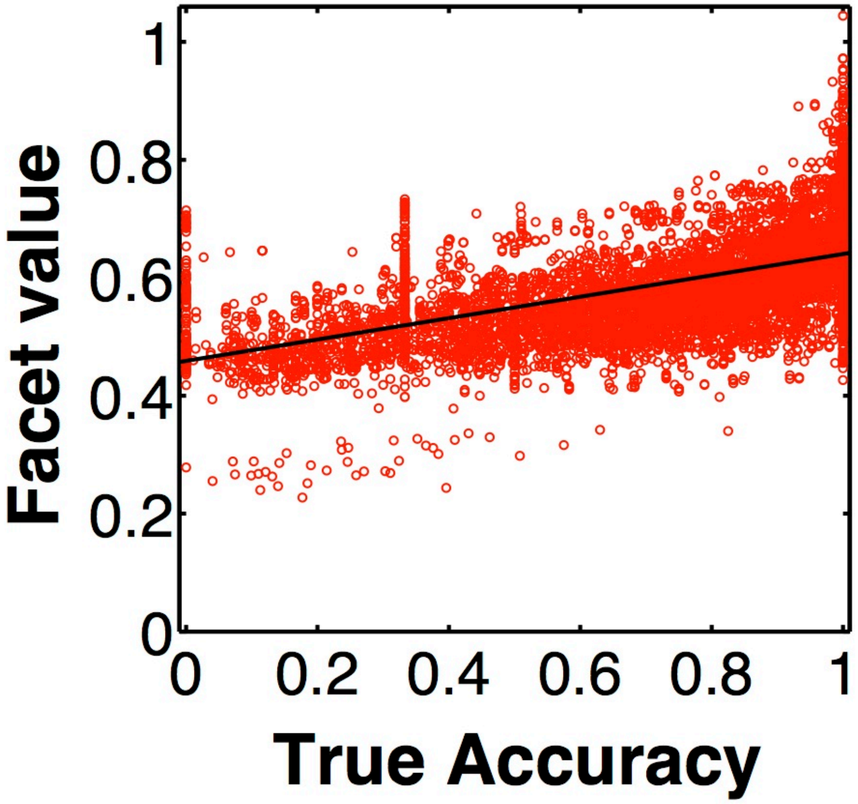
- accuracy values of the examples, or
- accuracy differences on pairs of examples.

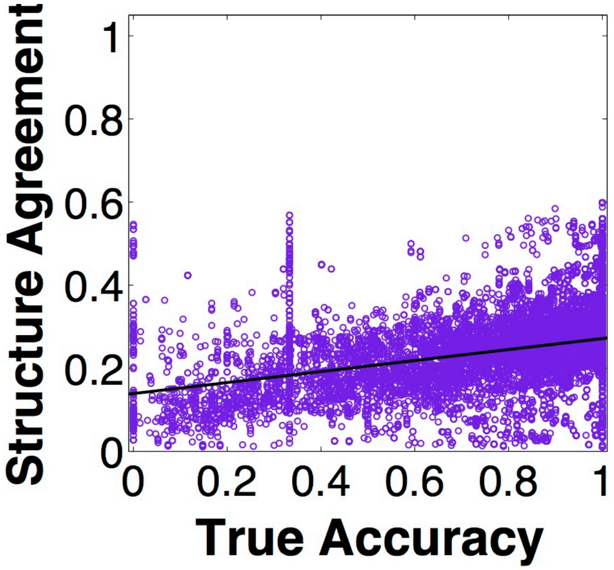


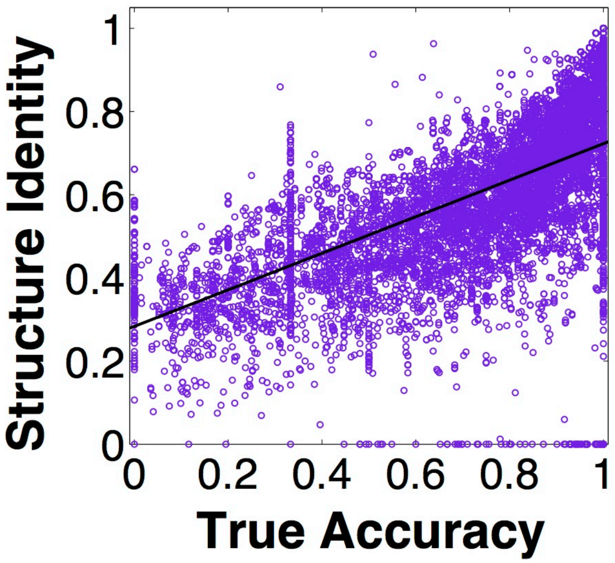
Experimental results: Advisor estimator

Best features trend well with accuracy.





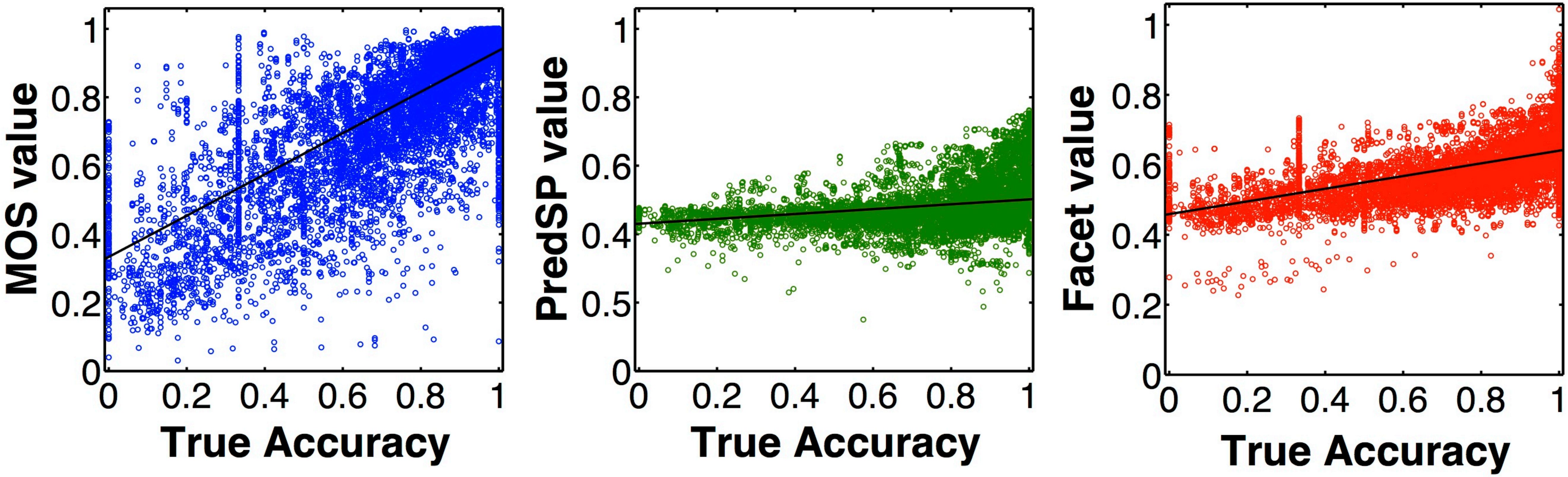




Facet estimator has better spread than its features.

Experimental results: Advisor estimator

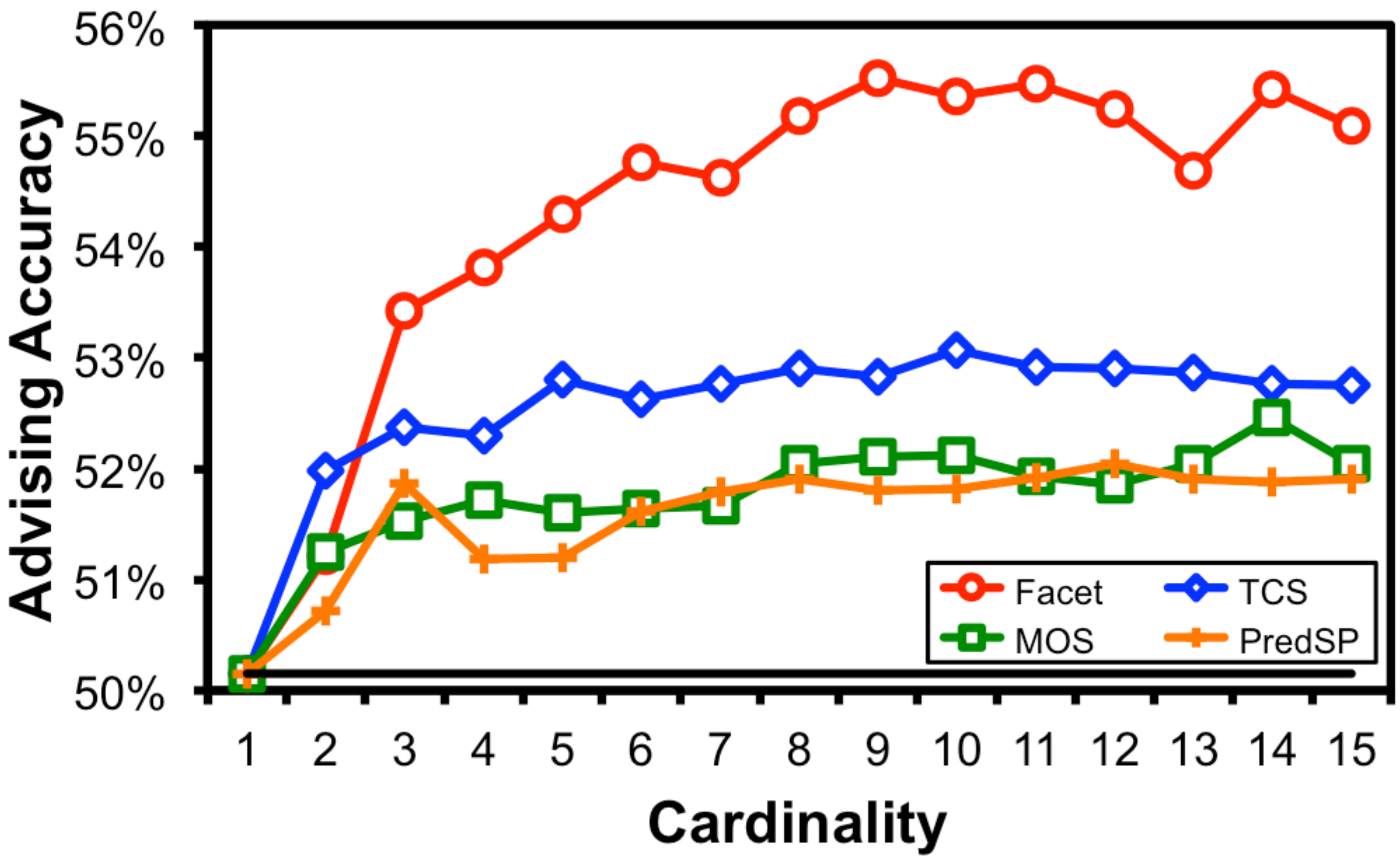
Known estimators display very different trends.



For parameter advising, an estimator needs to have good slope and spread.

Experimental results: Advisor estimator

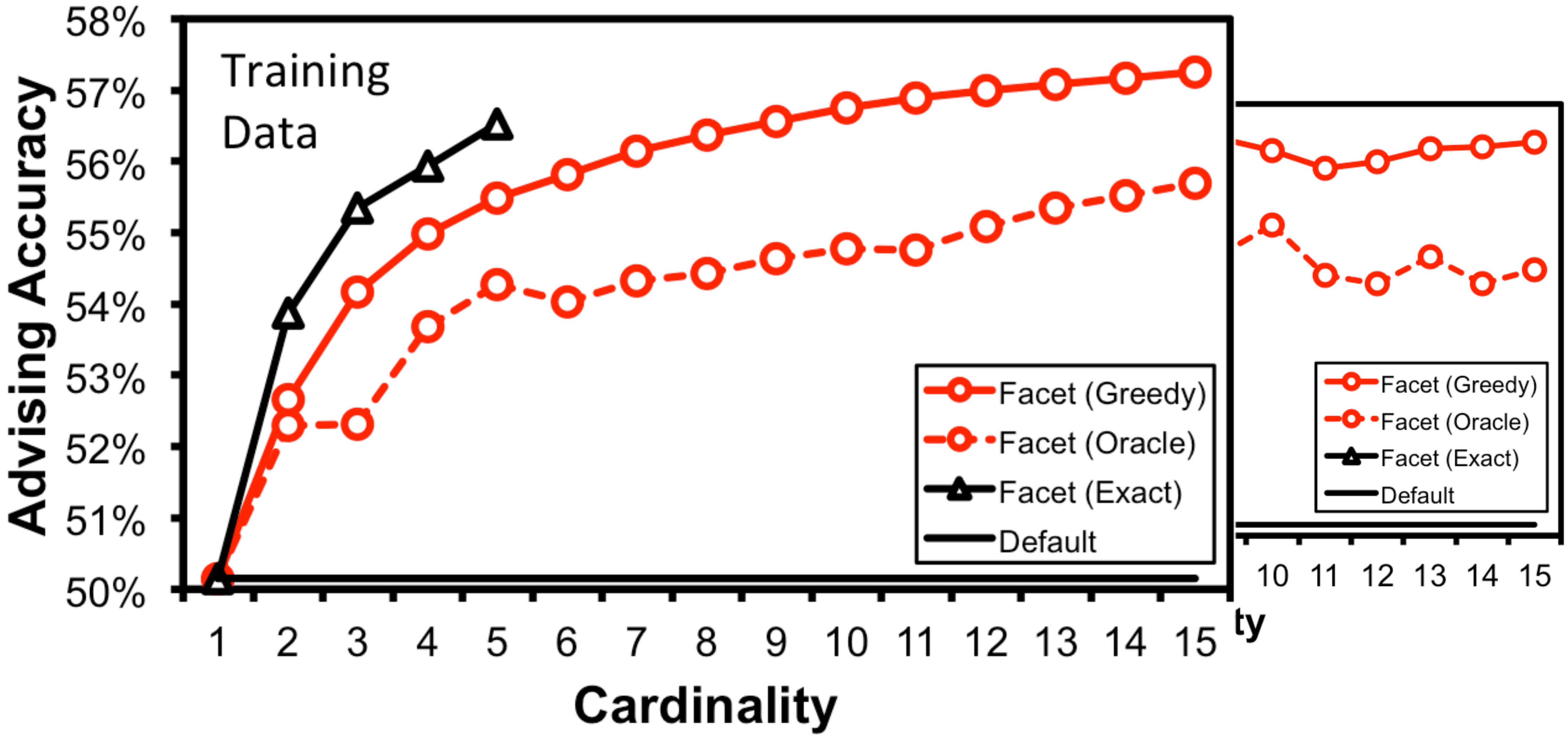
Advising accuracy of various accuracy estimators



As the cardinality of P increases, Facet accuracy increases.

Experimental results: Advisor sets

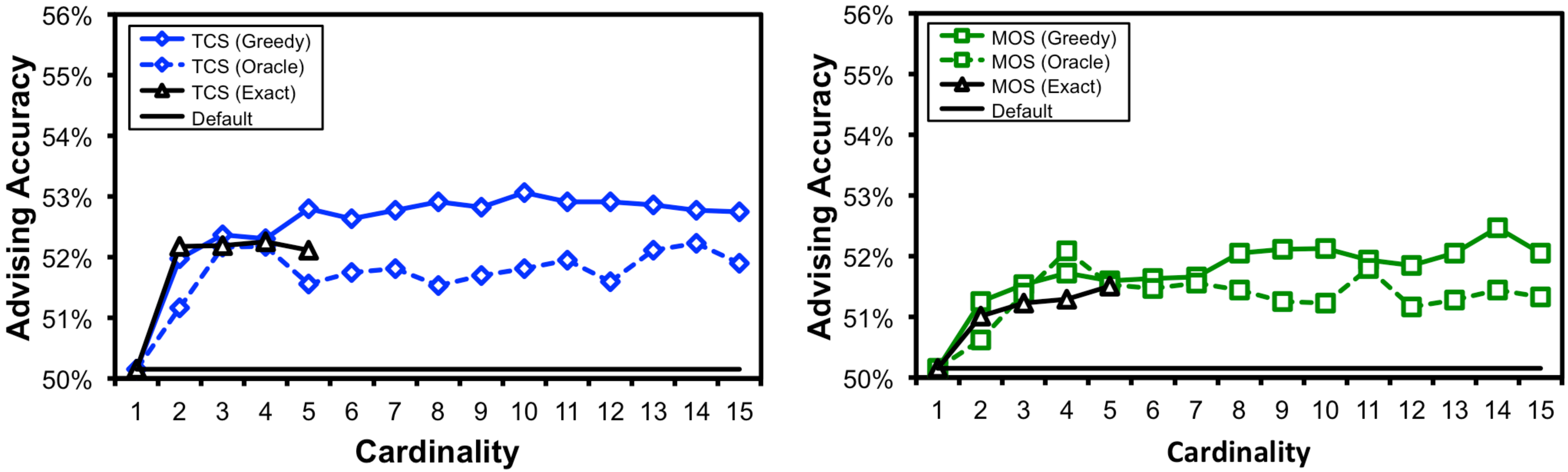
Advising accuracy on oracle, exact and greedy parameter sets



The greedy set is essentially as good as the optimal exact set.

Experimental results: Advisor sets

Advising accuracy on oracle, exact and greedy parameter sets



Finding advisor sets improves accuracy of other estimators.

Summary

Our current work has made the following contributions:

- New estimator Facet that is significantly more accurate for parameter advising
- Problem formulations for learning an advisor that are NP-complete
- Difference-fitting technique for estimator coefficients that is close to optimal
- Approximation algorithm for advisor sets that is close to optimal

Further research

- Develop a core column predictor for feature functions
- Extend the estimator from protein to DNA alignments
- Expand the definition of a parameter choice to include the aligner.

Thank you

People

- Travis Wheeler
  HHMI Janelia Farm
- Vladimir Filkov
  UC Davis

Thesis Committee Members:

- Alon Efrat, CS
- Stephen Kobourov, CS
- Mike Sanderson, EEB

Come see my poster

Today: 31

Sunday: N25

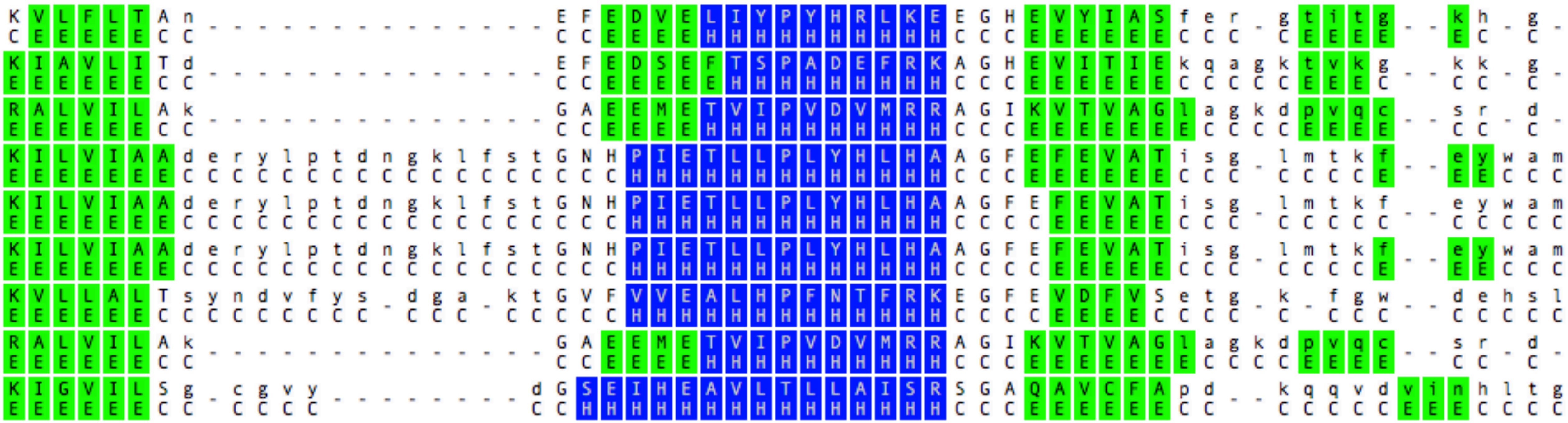
Funding

- NSF Grant IIS-1217886
- ISCB Student Council Travel Grant

Feature functions

There are three types of protein secondary structure

- α-helix,
- β-strand,
- coil.



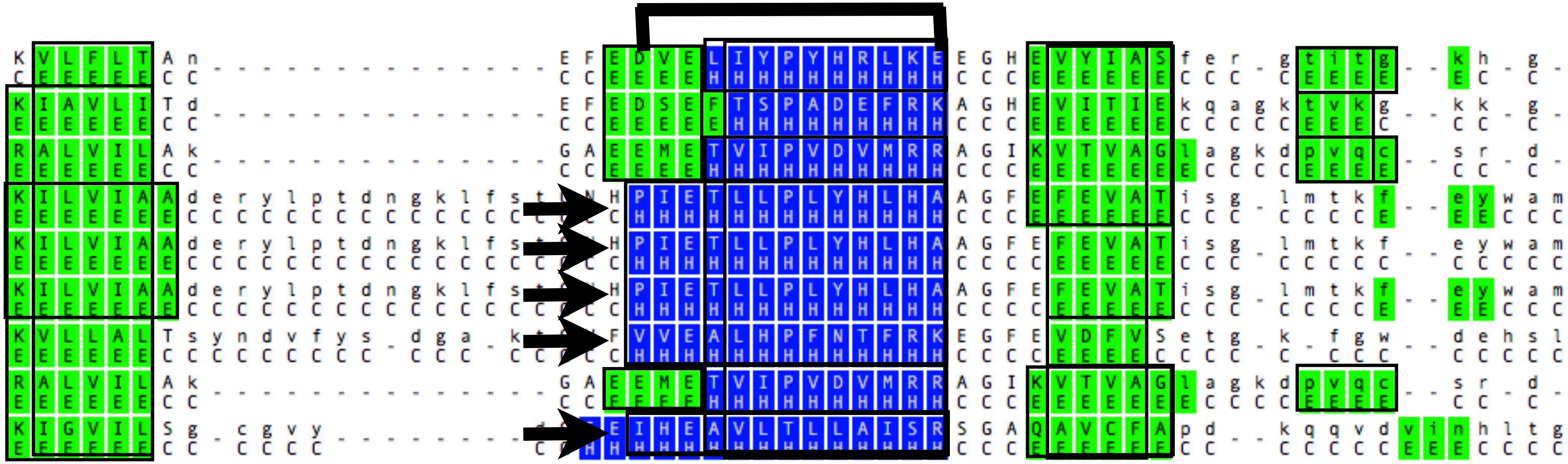
```
K V L F L T A n - - - - - - - - - - - - - - E F E D V E L I Y P Y H R L K E E G H E V Y I A S f e r - g t i t g - - k h - g -
C E E E E E C C - - - - - - - - - - - - - - C C E E E E H H H H H H H H H H C C C E E E E E E C C C - C E E E E - - E C - C -

K I 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 -
E E E E E E C C - - - - - - - - - - - - - - C C E E E E E H H H H H H H H H C C C E E E E E E C C C C C E E E C - - C C - C -

R A L V I L A k - - - - - - - - - - - - - - G A 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 - - s r - d -
E E E E E E C C - - - - - - - - - - - - - - C C E E E E H H H H H H H H H H C C C E E E E E E E C C C C E E E E - - C C - C -

K I L V 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 Y H L H A A G F E F E V A T i s g - l m t k f - - e y w a m
E E E E E E E C C C C C C C C C C C C C C C C C C H H H H H H H H H H H H H C C C E E E E E E C C C - C C C C E - - E E C C C

K I L V 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 Y H L H A A G F E F E V A T i s g - l m t k f - - e y w a m
E E E E E E E C C C C C C C C C C C C C C C C C C H H H H H H H H H H H H H C C C C E E E E E C C C - C C C C C - - C C C C C

K I L V 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 Y H L H A A G F E F E V A T i s g - l m t k f - - e y w a m
E E E E E E E C C C C C C C C C C C C C C C C C C H H H H H H H H H H H H H C C C C E E E E E C C C - C C C C E - - E E C C C

K 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 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
E E E E E E C C C C C C C C C - C C C - C C C C C H H H H H H H H H H H H H C C C C E E E E C C C C - C - C C C - - C C C C C

R A L V I L A k - - - - - - - - - - - - - - G A 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 - - s r - d -
E E E E E E C C - - - - - - - - - - - - - - C C E E E E H H H H H H H H H H C C C E E E E E E E C C C C E E E E - - C C - C -

K I G V I 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
E E E E E E C C - C C C C - - - - - - - - C C H H H H H H H H H H H H H H H C C C E E E E E E C C - - C C C C C E E E C C C C
```

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

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 positions in B.



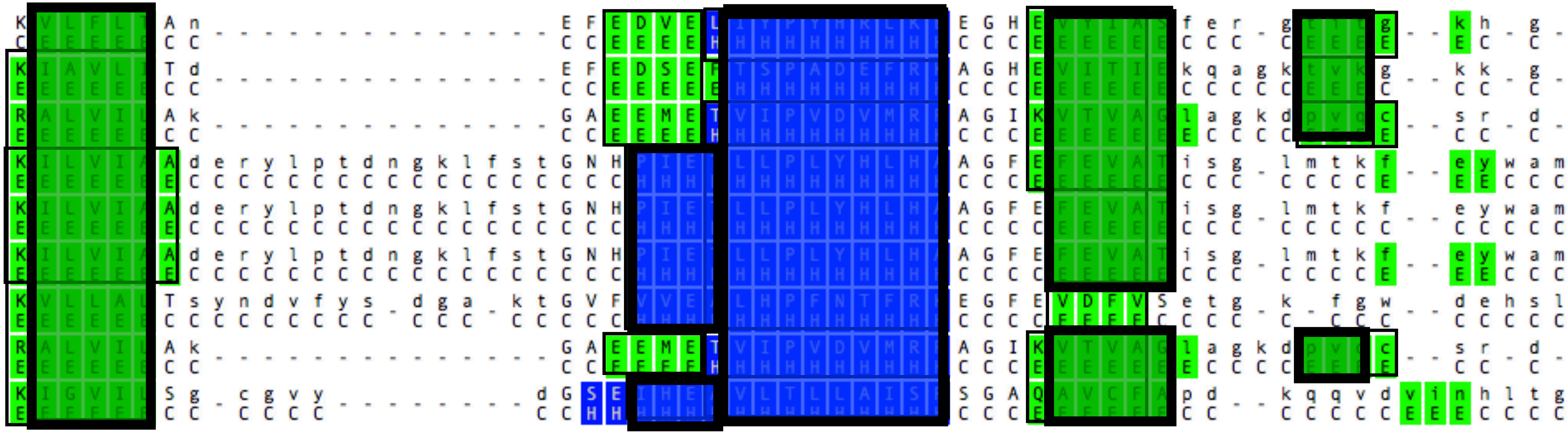
```
K V L F L T A n - - - - - - - - - - - - - - E F E D V E L I Y P Y H R L K E E G H E V Y I A S f e r - g t i t g - - k h - g -
C E E E E E C C                             C C E E E E H H H H H H H H H H C C C E E E E E E C C C   C E E E E     E C   C
K I A V L I T d - - - - - - - - - - - - - - E F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

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.



Secondary structure blockiness

The blockiness score of an alignment is

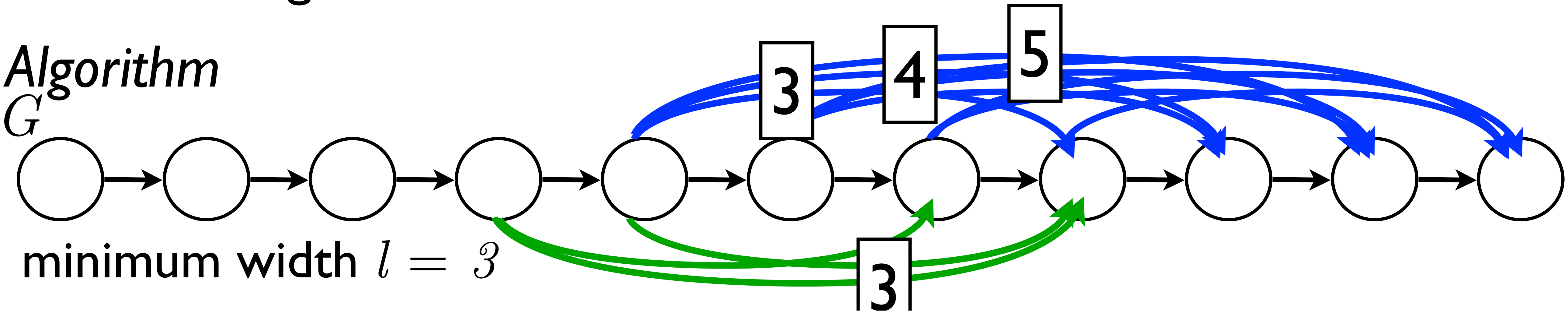
- the maximum value of *any* packing P of an alignment A
- normalized by the total number of substitutions in the alignment

Secondary Structure Blockiness

Theorem (Evaluating Blockiness)

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

Algorithm
G

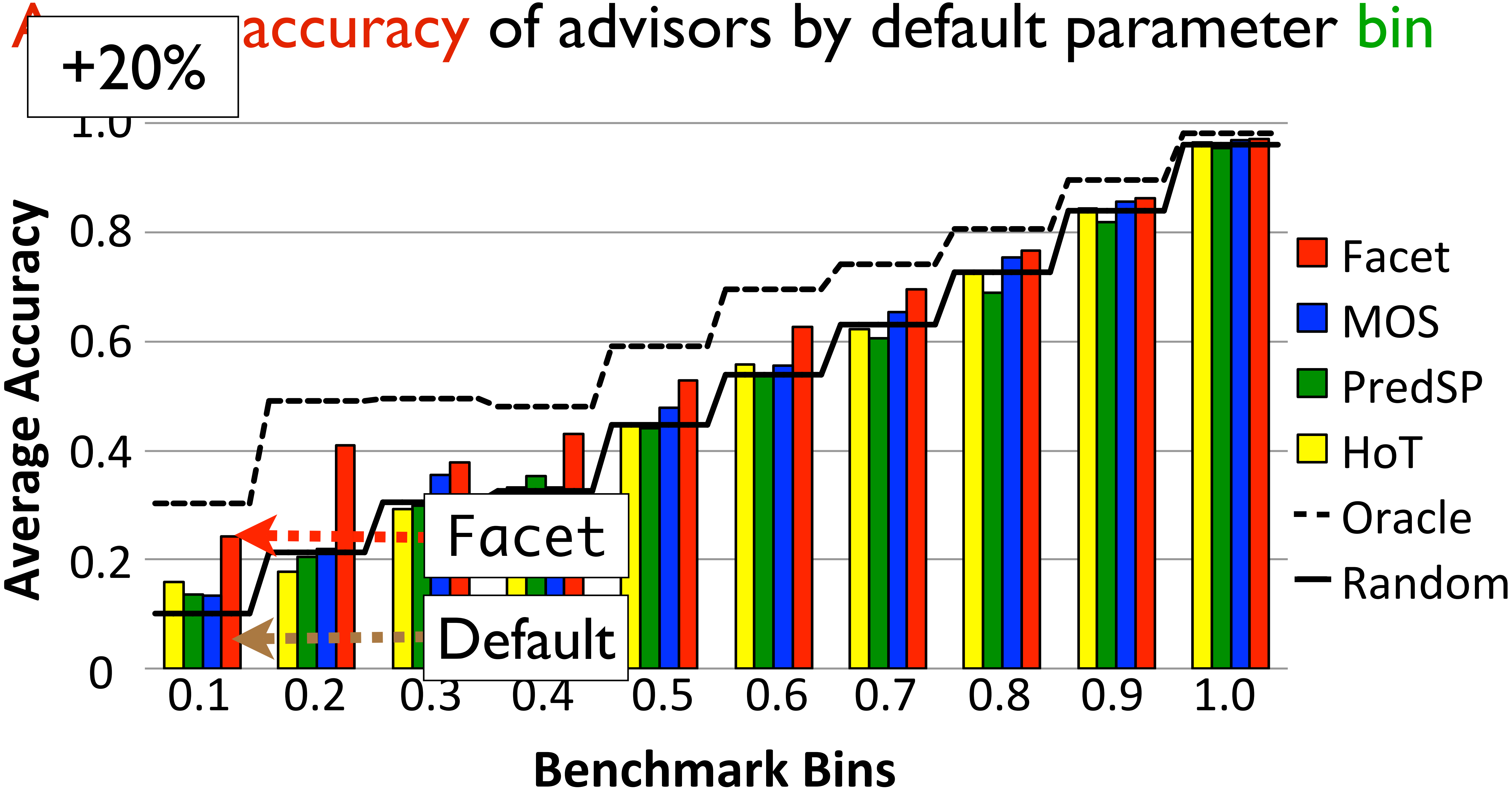


minimum width $l = 3$

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

Results

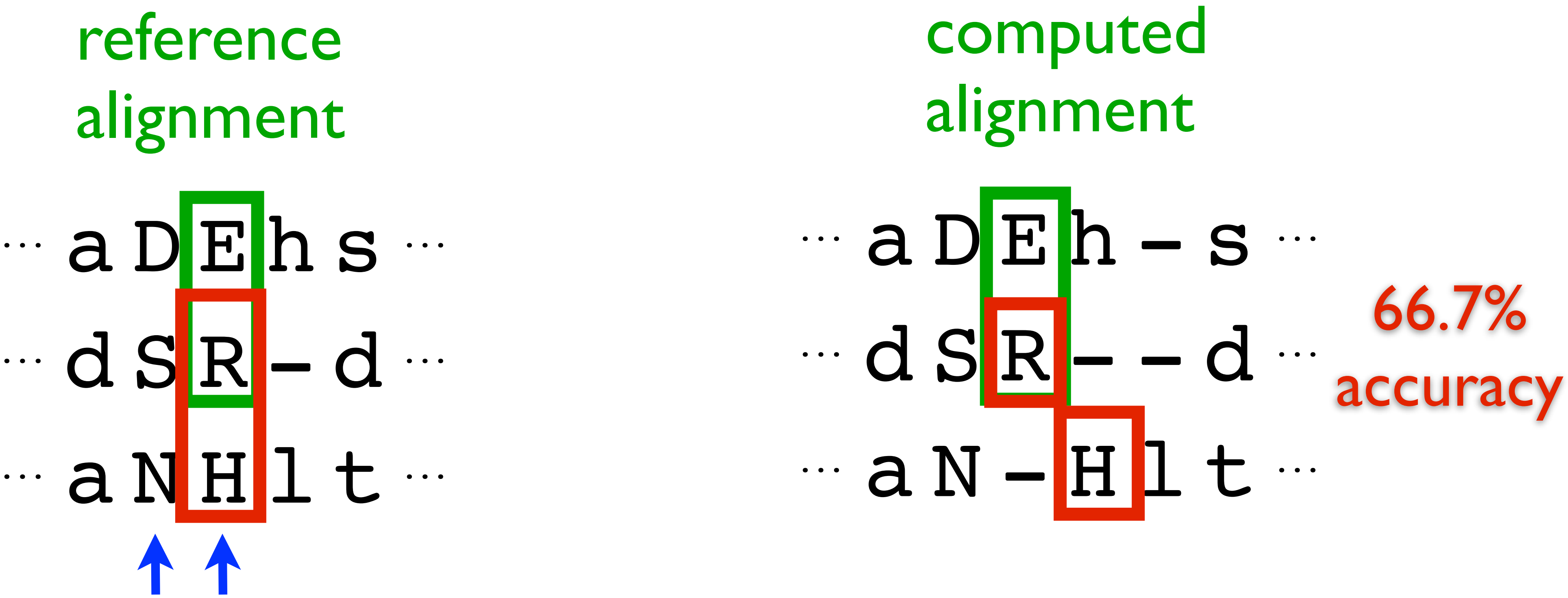
A accuracy of advisors by default parameter bin



In all bins, Facet outperforms all estimators.

Motivation

Alignment accuracy is measured with respect to a reference alignment.



reference alignment

```
··· a D E h s ···
··· d S R - d ···
··· a N H l t ···
```

computed alignment

```
··· a D E h - s ···
··· d S R - - d ···
··· a N - H l t ···
```

66.7% accuracy

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

Contributions

Our approach Facet (“Feature-based ACcuracy EsTimator”)

- estimates accuracy by a polynomial on the features,
- efficiently learns the polynomial coefficients from examples,
- uses novel features that are fast to evaluate,
- utilizes an optimal feature subset.

Applied to parameter advising, Facet:

- finds an optimal parameter set of a given cardinality,
- outperforms other estimators in accuracy across the full range of benchmarks,
- boosts aligner accuracy on hard benchmarks by 20% over the best default parameter choice.

Optimal Advisor

The input is

- cardinality bound k,
- weights w_i on the benchmarks,
- accuracies a_{ij} of the alternate alignments,
- feature vectors F_{ij} for the alternate alignments, and
- an error tolerance ε,

Output

- set $P \subseteq \{1, \ldots, m\}$ of parameter choices where $|P| \leq k$, and
- estimator coefficients $c = (c_1, \ldots, c_l) \in \mathcal{Q}$

Learning the estimator

Difference-fitting tries to find a monotonic estimator that matches positive differences in true accuracy.

$$c^* := \underset{c\,\in\,\mathcal{R}^t}{\operatorname{argmin}} \sum_{(A,B)\,\in\,\mathcal{P}} w_{AB}\left(\max\left\{\big(F(B)-F(A)\big)-\big(E_c(B)-E_c(A)\big),0\right\}\right)^p$$

all possible coefficients

all important pairs of examples

true accuracy difference

estimated difference

controls influence of large errors

only penalize underestimating differences