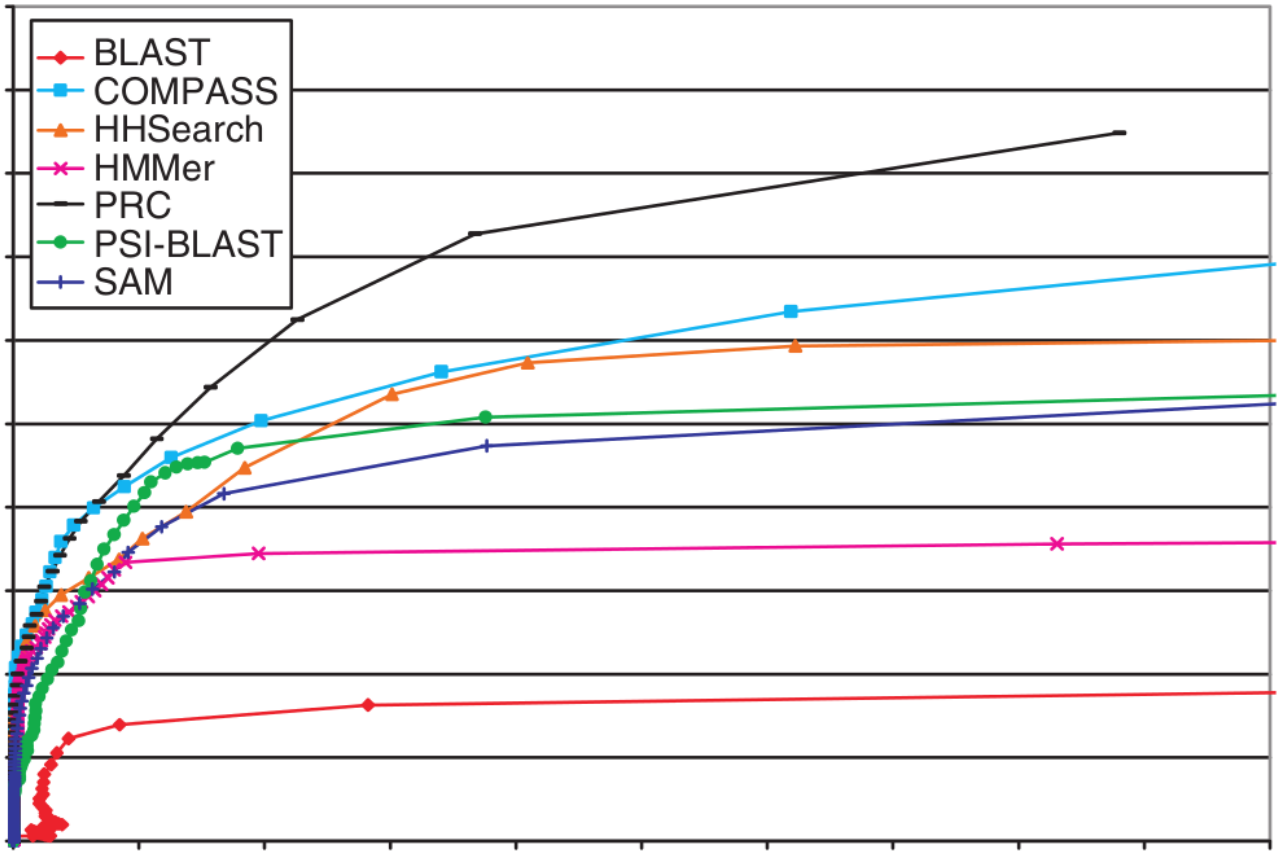
Lecture 5:
Homology detection and Benchmarking



17 Sept 2012

Fundamentals of Bioinformatics

Main points

- Benchmarking: common and critical task in bioinformatics tool development
- Needle in a haystack
- Scoring depends on (availability) of gold standard data
- Different ways to score
  - Measures
  - Plots
- Case: (Remote) Homolog Detection

Today's lecture

- Intro on benchmarking
- Work on questions in 'expert' groups
- Rearrange in 'jigsaw' groups; explain answers
- Wrapping up and additional/advanced benchmarking

Needle in a Haystack

- Main question: How good is your BLAST-hit?
  - Is it a real homolog or not?
- Depends on similarity
  - Close homolog (low e-value)
    → unlikely to be a random match
  - Distant homolog (high e-value)
    → many more options, so high *a priori* random chance
- Scoring in benchmark needs to balance

Scoring

	database entry:	*hit in same category as query*	*hit in different category from query*	*hit not in database*
data:	**conclusion:**	true homolog	no homolog	unknown
BLAST hit (e-value better than threshold)	*putative homolog*	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)	UNKNOWN
no hit (e-value worse than threshold)	*putative non-homolog*	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)	UNKNOWN

- True positive: correct prediction
- False positive: wrongly predicted
- True negative: correctly not predicted
- False negative: missed prediction

Some easy things to remember

- All positives (P): TP + FN
- All negatives (N): TN + FP
- All (positive) predictions: TP + FP

- And, just to remind you:
  - True positive: correct prediction
  - False positive: wrongly predicted
  - True negative: correctly not predicted
  - False negative: missed prediction

An example!

	Female	Male
short		
long		

- Predictions:
  - 'A female person has long hair'
  - 'A person with short hair is male'
- How good are these predictions?

Different measures

Fractions of all positives (TP+FN) or negatives (TN+FP):

- Sensitivity, Coverage, Recall, True positive rate (TPR)

  $Sn = TP / (TP+FN)$

- Specificity (1–Error) $Sp = TN / (TN+FP)$
- Error (1–Specificity), Fall-Out, False positive rate (FPR)

  $Err = FP / (TN+FP)$

Fractions of all (positive) predictions (TP+FP):

- Positive predictive value, precision (1–EPQ)

  $PPV = TP / (TP+FP)$

- Error per Query (1–PPV) $EPQ = FP / (TP+FP)$

An example!

	Female	Male
short		
long		

- Predictions:
  - 'A female person has long hair'
  - 'A person with short hair is male'
- calculate the following:
  - Sensitivity
  - Specificity
  - Error
  - Precision
  - Error per Query

Tradeoff

- It is easy to predict all known cases!
  - How to do that?
  - Which measure do you optimize?

- It is easy not to make any errors!
  - How to do that?
  - Which measure do you optimize?

- In both cases, what happens to the other measure(s)?

Some more measures:

- Accuracy = (TP + TN) / (TP+FN+TN+FP)
- Balanced accuracy
  = (Sensitivity + Specificity) / 2
  = ½ TP / (TP+FN) + ½ TN / (TN+FP)
- Matthews correlation coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- (there are many other possible measures)

An example!

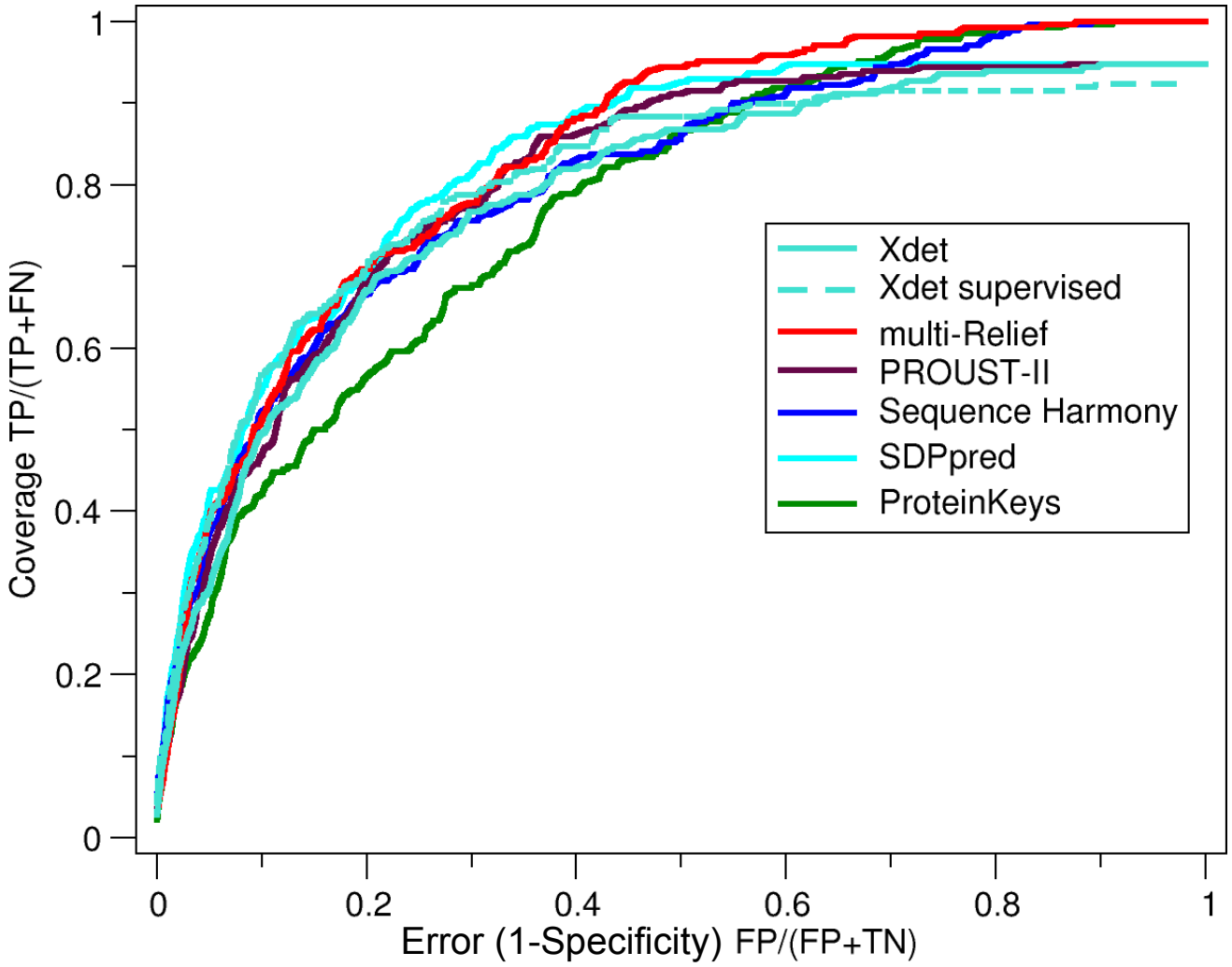
	Female	Male
short		
long		

- Predictions:
  - 'A female has long hair'
  - 'A person with short hair is male'
- How could we optimize Accuracy (Sensitivity + Specificity) for these predictions?

Different plots

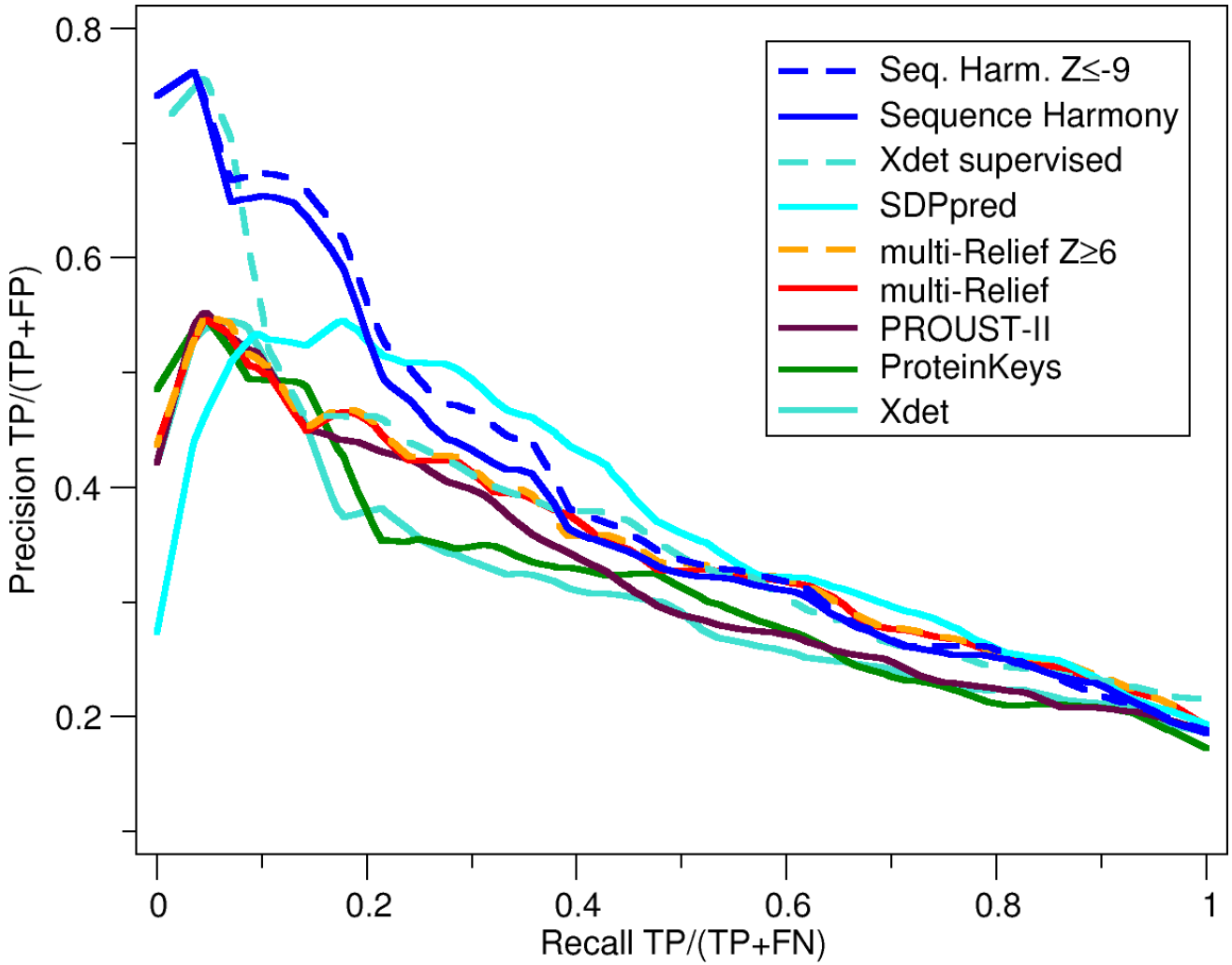
- Reciever-Operator Characteristics (ROC):
  - Coverage TP/(TP+FN) *vs.* Specificity FP/(FP+TN)
- Precision/Recall:
  - Precision FP/(TP+TN) *vs.* Recall TP/(TP+FN)
- Coverage/EPQ:
  - Coverage TP/(TP+FN) *vs.* Error per Query FP/(TP+FP)

Receiver-Operator Characteristics (ROC) Plot (methods of specificity residue detection)



CENTRE FOR INTEGRATIVE BIOINFORMATICS VU

Example Precision/Recall (P/R) Plot (methods of specificity residue detection)

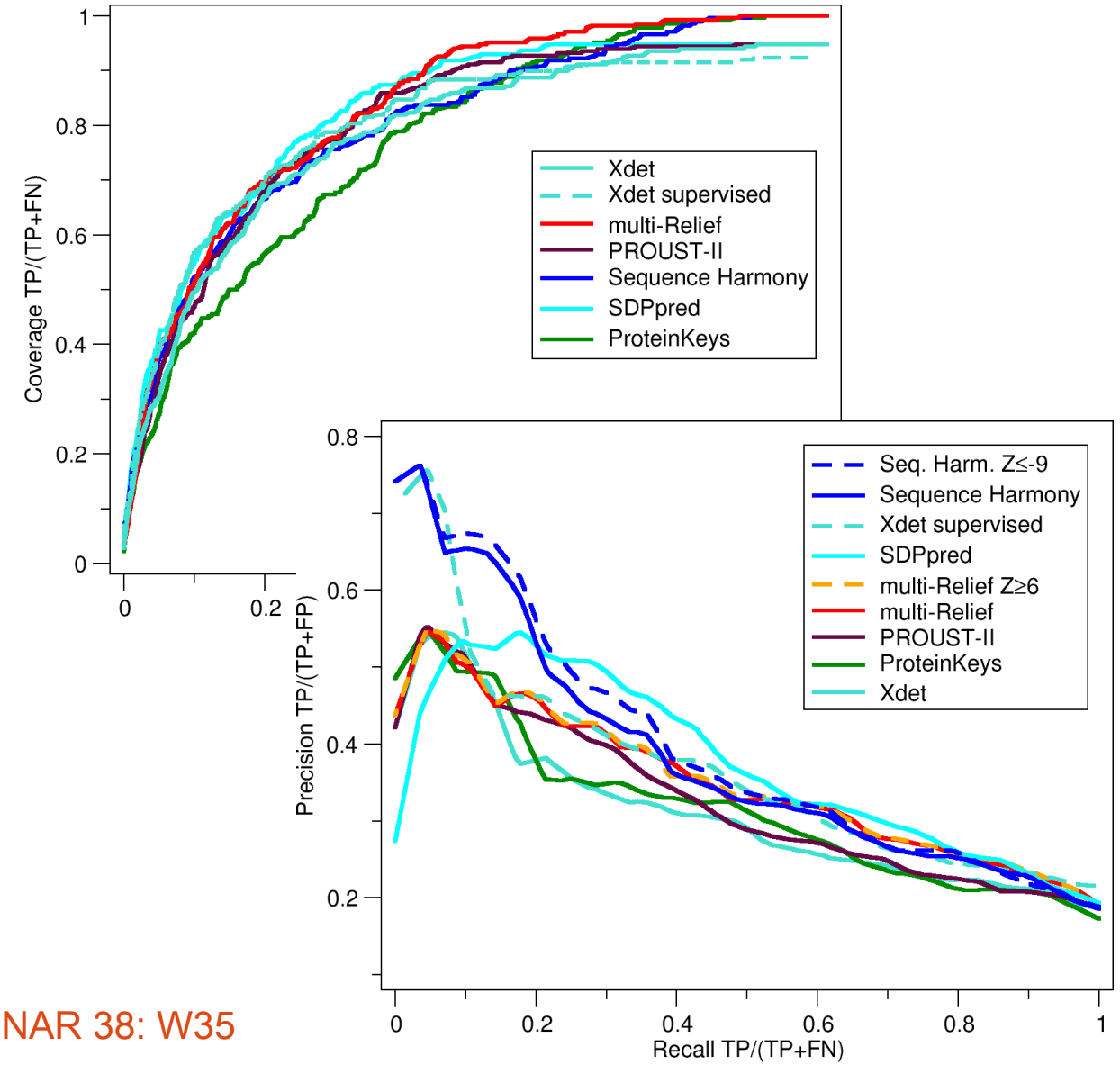


Brandt, Feenstra, Heringa (2010) NAR 38: W35

CENTRE FOR INTEGRATIVE BIOINFORMATICS VU

Different representations: ROC *vs.* P/R

- How do the axes relate between a ROC and a P/R plot?
- What question is implicitly asked in each of the plots?

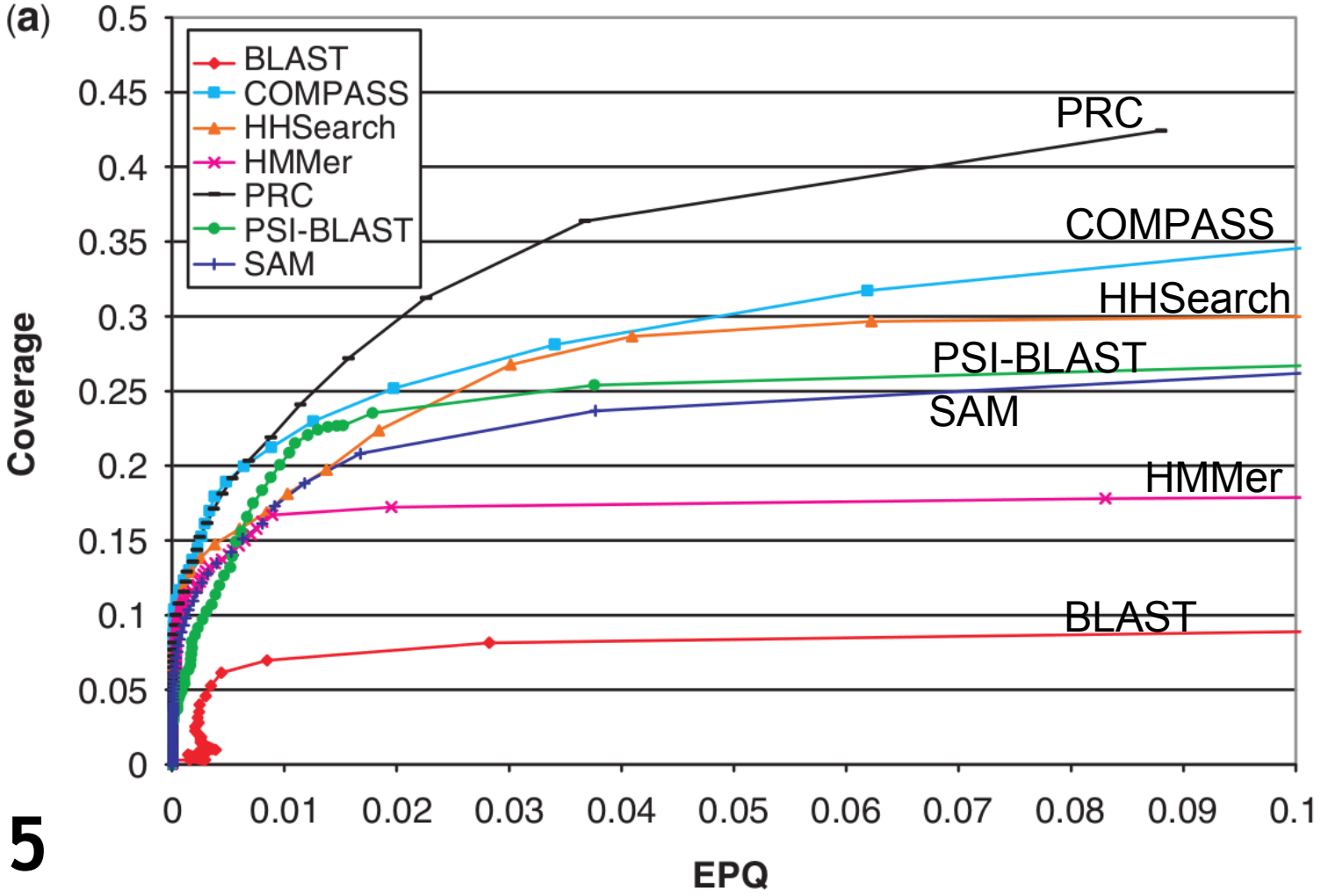


Brandt, Feenstra, Heringa (2010) NAR 38: W35

Today's lecture

- Intro on benchmarking
- Work on questions in 'expert' groups
  - 5 minutes to make 8 groups, ~7 persons each
  - 10 minutes to work on question
- Rearrange in 'jigsaw' groups; explain answers
- Wrapping up and additional/advanced benchmarking

Homology detection methods



nr35

Reid et al. Bioinf v23 p2353 (2007) Fig 3a

Paper questions:

1. Why do they use Coverage vs. Error per Query and not a ROC plot?
2. How do the different methods compare, and can you divide them into different performing groups?
3. Which reference database did they use for benchmarking, and is this a wise choice?
4. Would you expect to same performance if you would run these methods on a set of trans-membrane proteins?

Reid, Yeats & Orengo Bioinformatics 23 2353-60 (2007)

"Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone"

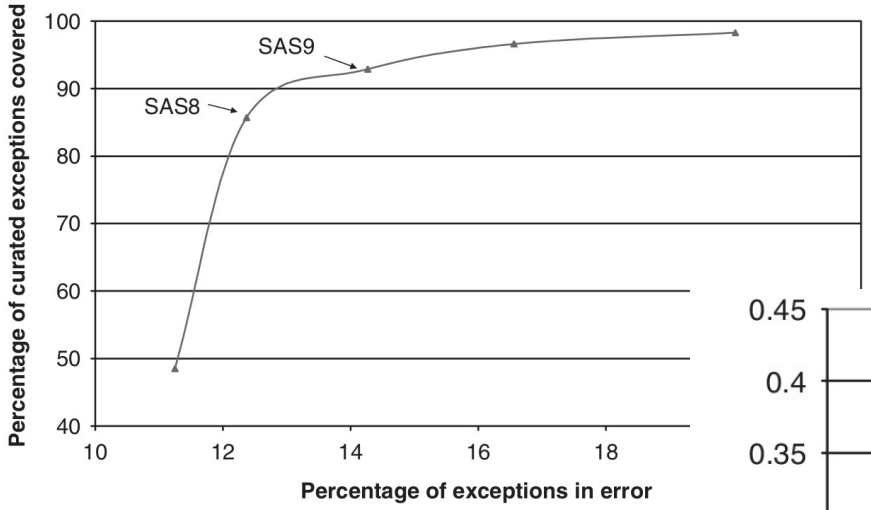
Today's lecture

- Intro on benchmarking
- Work on questions in 'expert' groups
- Rearrange in 'jigsaw' groups; explain answers
  - After break, 5 minutes to make groups
  - Max two persons per question in one group!
  - 20 minutes to work on questions (5 min per question)
- Wrapping up and additional/advanced benchmarking

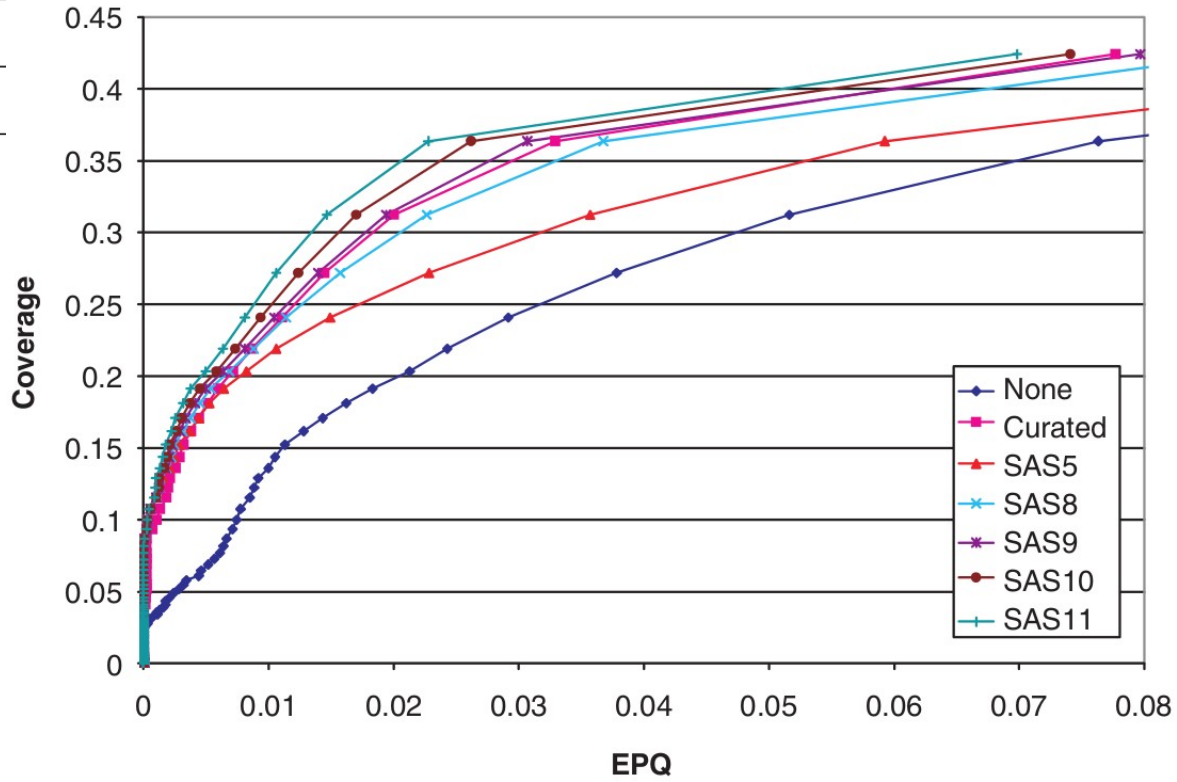
Today's lecture

- Intro on benchmarking
- Work on questions in 'expert' groups
- Rearrange in 'jigsaw' groups; explain answers
- Wrapping up and additional/advanced benchmarking

Importance of reference data

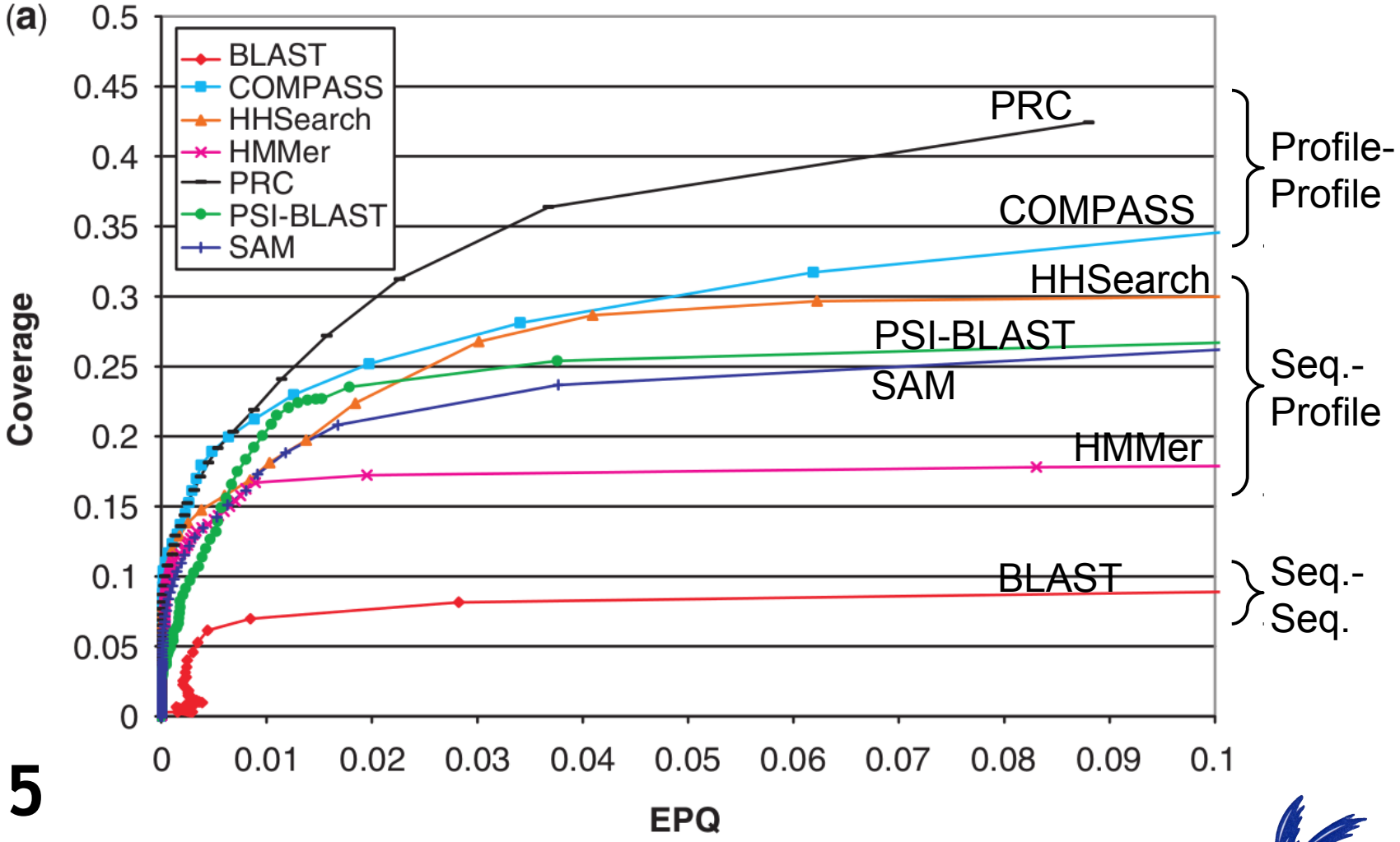


- CATH:
  Protein structure classification database
- Heuristic rule for exceptions
  (based on SAS score)



Reid et al. Bioinf v23 p2353 (2007) Fig 2

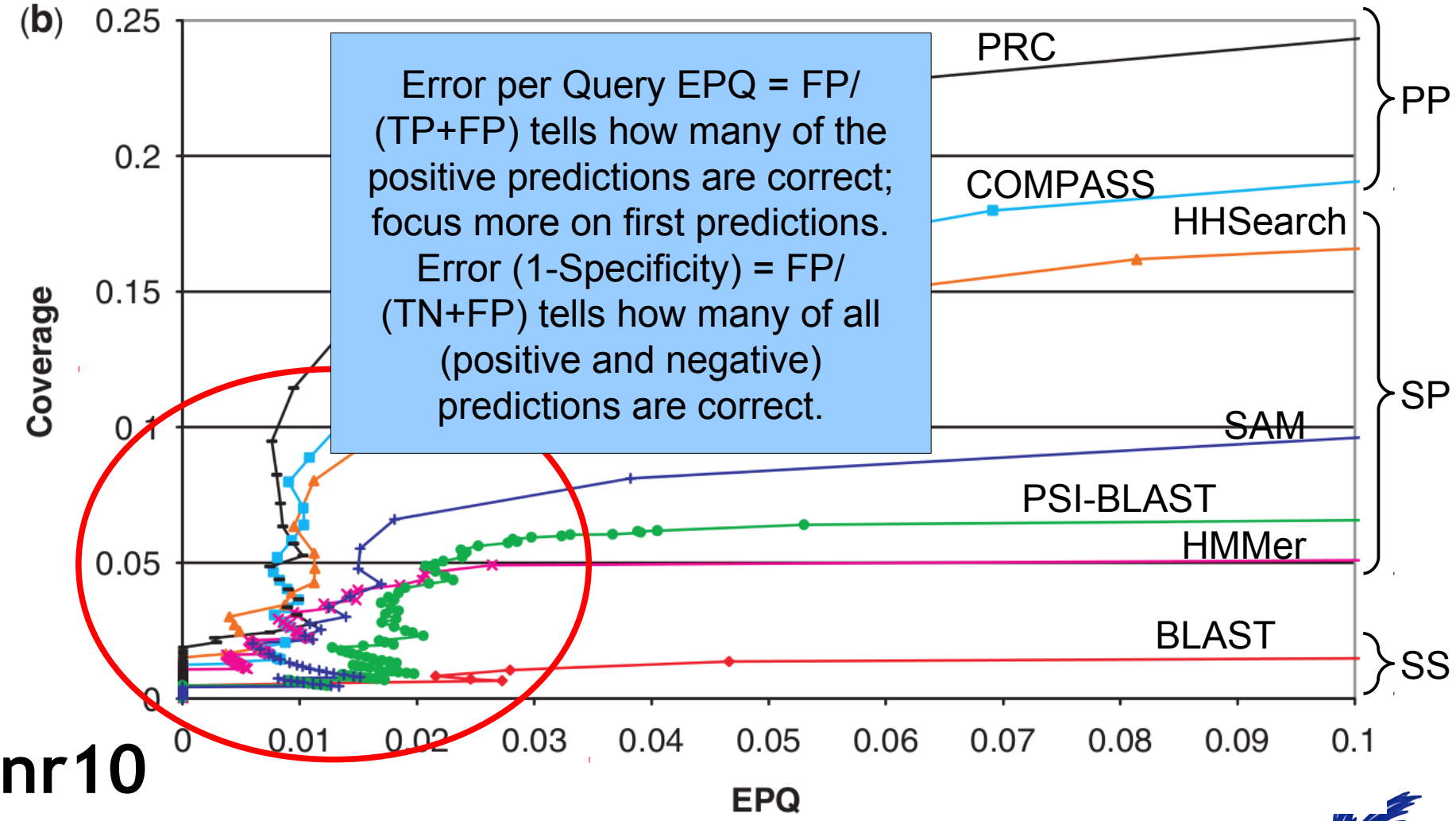
Homology detection methods: Better coverage for profile-profile methods



nr35

Reid et al. Bioinf v23 p2353 (2007) Fig 3a

Homology detection methods for *nr10*: Larger differences at low sequence identity



nr10

Reid et al. Bioinf v23 p2353 (2007) Fig 3b

CENTRE FOR INTEGRATIVE BIOINFORMATICS VU

Main points

- Benchmarking: common and critical task in bioinformatics tool development
- Needle in a haystack
  - balance sensitivity and specificity
- Scoring depends on (availability) of gold standard data
- Different ways to score
  - Measures: Sensitivity, Specificity, Error, Positive predictive value, Error per Query
  - Plots: ROC, P/R, Coverage/EPQ
- Case: (Remote) Homolog Detection
  - Strength of Profile-Profile scoring

Lecture 5: Homology detection and Benchmarking

Fundamentals of Bioinformatics

Benchmarking
17 Sept 2012