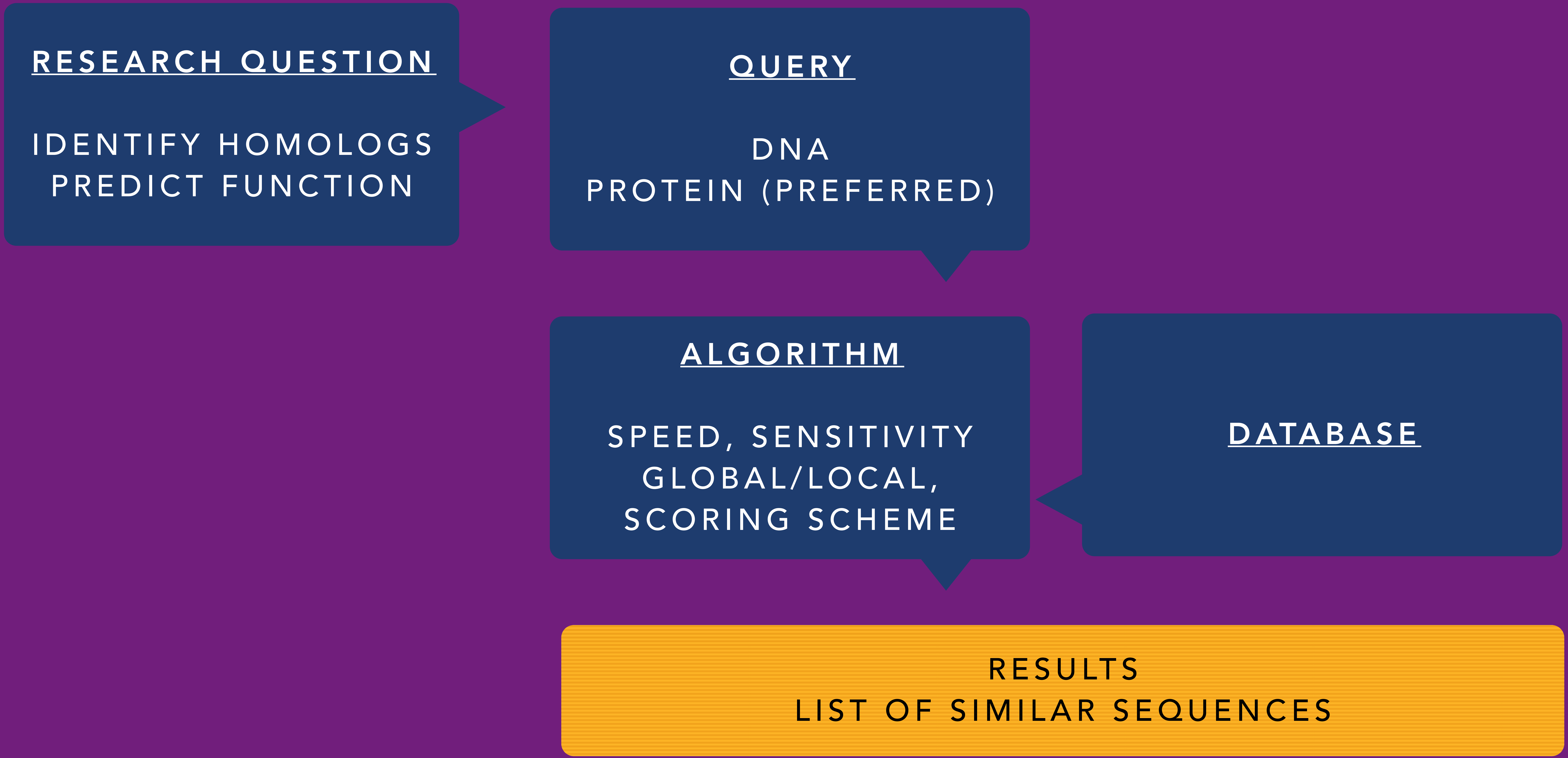
DATABASE SEARCHING



RESEARCH QUESTION

IDENTIFY HOMOLOGS
PREDICT FUNCTION

QUERY

DNA
PROTEIN (PREFERRED)

ALGORITHM

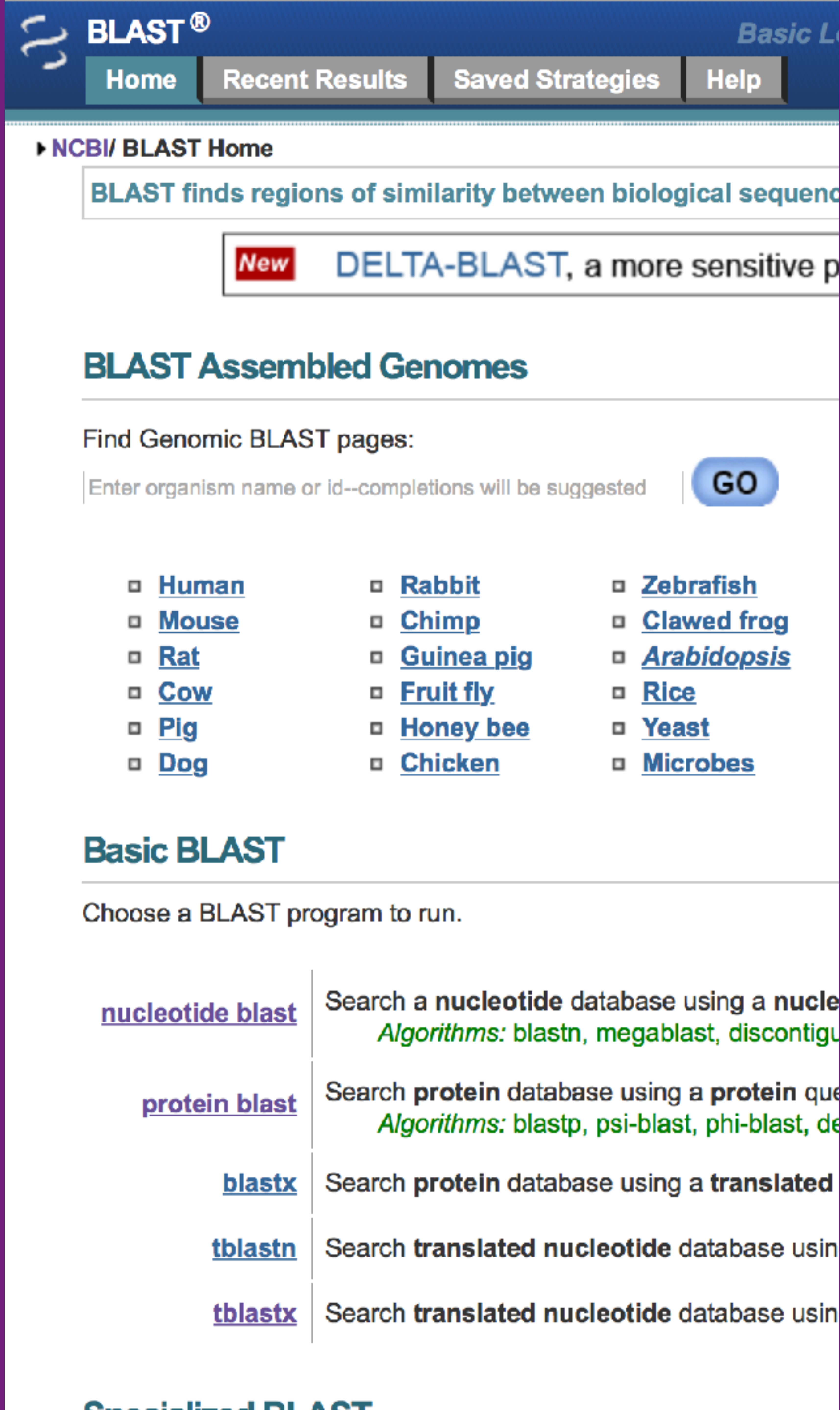
SPEED, SENSITIVITY
GLOBAL/LOCAL,
SCORING SCHEME

DATABASE

RESULTS
LIST OF SIMILAR SEQUENCES

DATABASE SEARCHING

- Interpret results
- Optionally change search strategy
  - Change algorithm
  - Change scoring scheme



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequenc

New DELTA-BLAST, a more sensitive p

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested GO

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

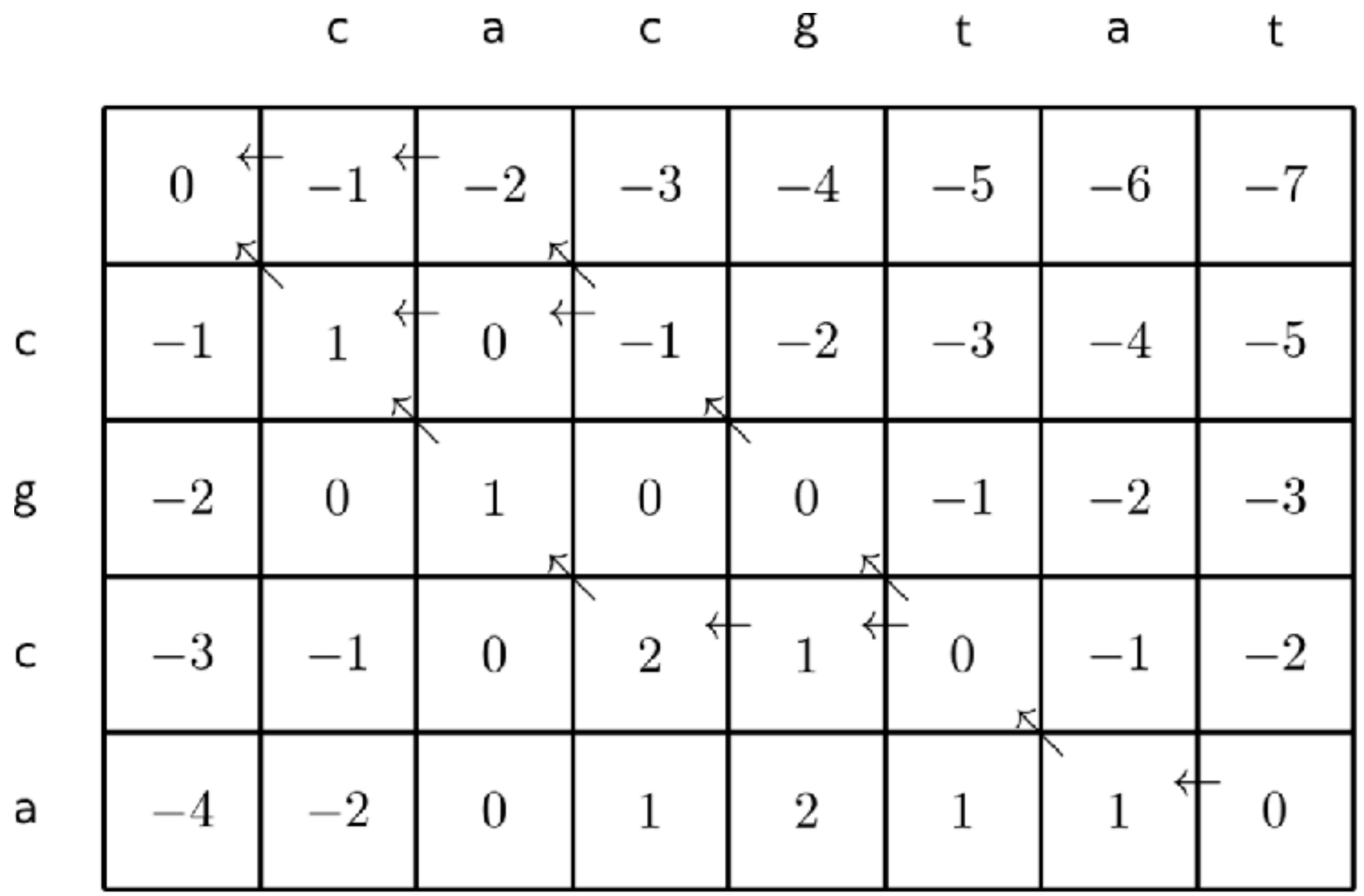
Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucle** *Algorithms:* blastn, megablast, discontigu
protein blast	Search **protein** database using a **protein** que *Algorithms:* blastp, psi-blast, phi-blast, de
blastx	Search **protein** database using a **translated**
tblastn	Search **translated nucleotide** database usin
tblastx	Search **translated nucleotide** database usin

Specialized BLAST

HEURISTIC SEQUENCE ALIGNMENT

- Optimal sequence alignment and statistics do not scale to database searches
  - Needleman-Wunch O(nm)
  - Smith-Waterman O(nm)

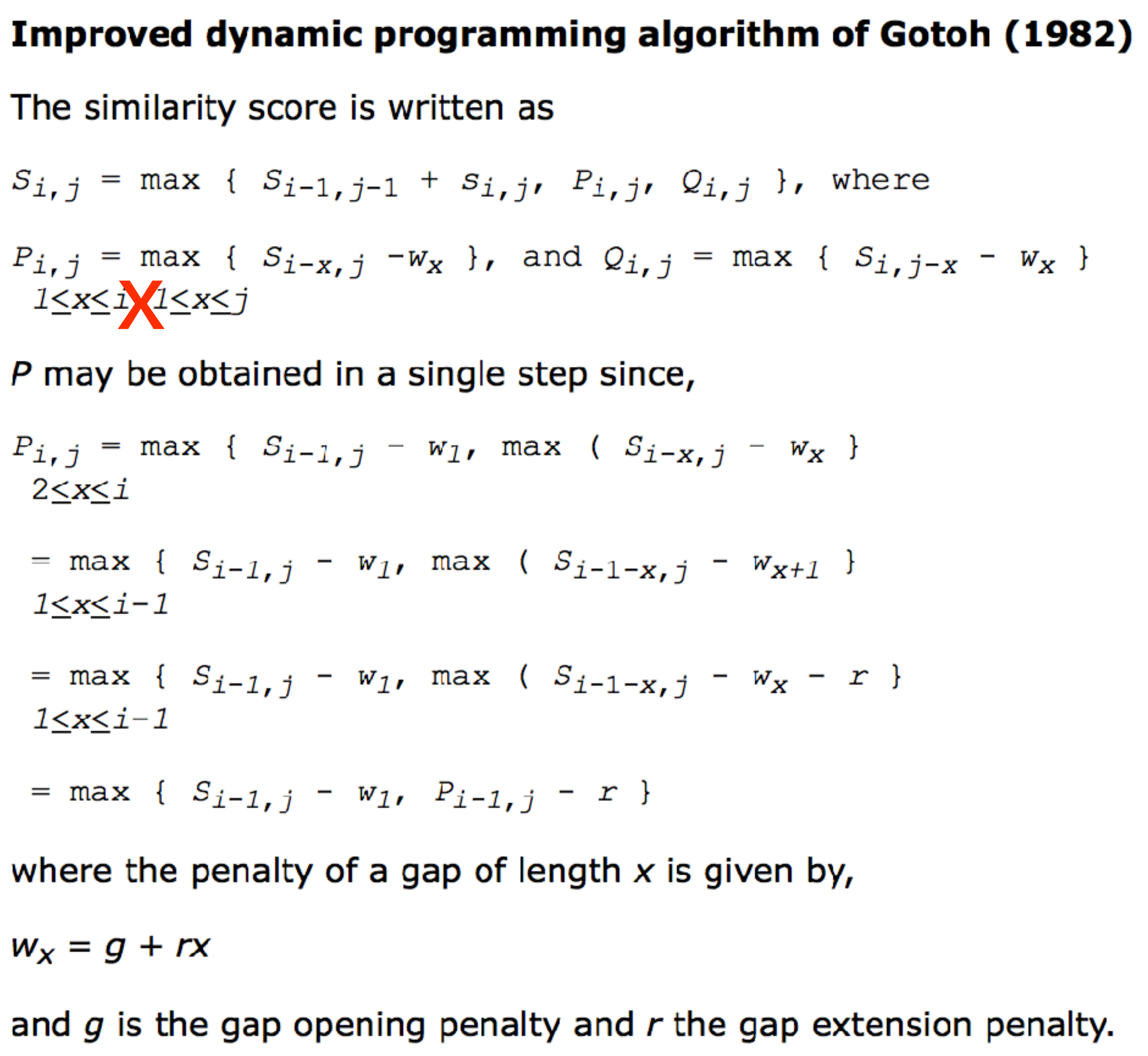


		c	a	c	g	t	a	t
	0	−1	−2	−3	−4	−5	−6	−7
c	−1	1	0	−1	−2	−3	−4	−5
g	−2	0	1	0	0	−1	−2	−3
c	−3	−1	0	2	1	0	−1	−2
a	−4	−2	0	1	2	1	1	0

HEURISTIC SEQUENCE ALIGNMENT

- Gotoh (1982) simplified the dynamic programming algorithm
  - Introduced the affine gap penalty
  - He reasoned that two of the terms that are maximized in the dynamic programming algorithm depend only on the values in the **current and previous row and column**

NO NEED TO CALCULATE THE ENTIRE MATRIX



Improved dynamic programming algorithm of Gotoh (1982)

The similarity score is written as

$S_{i,j} = \max \{ S_{i-1,j-1} + s_{i,j}, P_{i,j}, Q_{i,j} \}$, where

$P_{i,j} = \max \{ S_{i-x,j} - w_x \}$, and $Q_{i,j} = \max \{ S_{i,j-x} - w_x \}$

$1 \le x \le i$ X $1 \le x \le j$

P may be obtained in a single step since,

$P_{i,j} = \max \{ S_{i-1,j} - w_1, \max (S_{i-x,j} - w_x \}$

$2 \le x \le i$

$= \max \{ S_{i-1,j} - w_1, \max (S_{i-1-x,j} - w_{x+1} \}$

$1 \le x \le i-1$

$= \max \{ S_{i-1,j} - w_1, \max (S_{i-1-x,j} - w_x - r \}$

$1 \le x \le i-1$

$= \max \{ S_{i-1,j} - w_1, P_{i-1,j} - r \}$

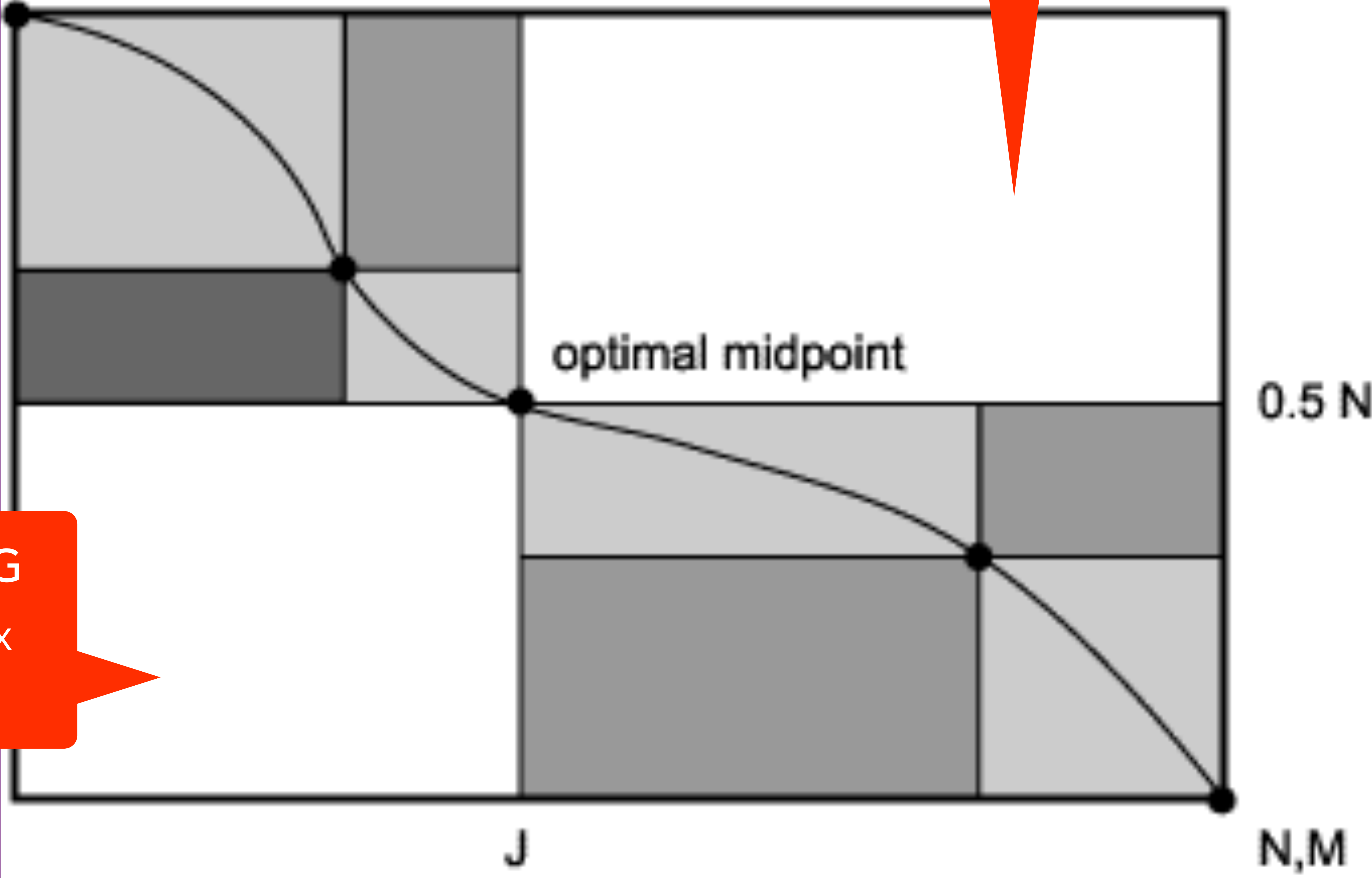
where the penalty of a gap of length *x* is given by,

$w_x = g + rx$

and *g* is the gap opening penalty and *r* the gap extension penalty.

HEURISTIC SEQUENCE ALIGNMENT

- Myers and Miller (1988)
    - Improved the algorithms so both global and local alignment require less time and sp
    - Start at beginning/end midpoint
    - Join alignments



NEVER GOING TO FIND S_{MAX} HERE

NEVER GOING TO FIND S_{MAX} HERE

~ .5*(MxN)

HEURISTIC SEQUENCE ALIGNMENT

- Why are heuristic approaches necessary?
  - Dynamic programming requires order **N^2L** computations where
    - **N** is size of the query sequence
    - **L** is the size of the database
- Given size of databases, more efficient methods needed

HEURISTIC SEQUENCE ALIGNMENT

- How are heuristic approaches applied?
  - Feedback from current result guides future analytical direction
  - Search a small fraction of the cells in possible search space
    - Still maintain all the high scoring alignments

HEURISTIC SEQUENCE ALIGNMENT

PART OF ITERATIVE STRATEGY

- Heuristic methods are **not guaranteed** to find the optimal solution
- Can be much faster
  - >>50X improvement in speed/memory usage

NECESSARY TO SEARCH AGAINST THE NCBI SEQUENCE DATABASE

HEURISTIC SEQUENCE ALIGNMENT

- 2 best known approaches
  - FASTA [Pearson & Lipman, 1988]
  - BLAST [Altschul et al., 1990]

TRADEOFFS OF USING THE HEURISTIC METHODS?

ACCURACY

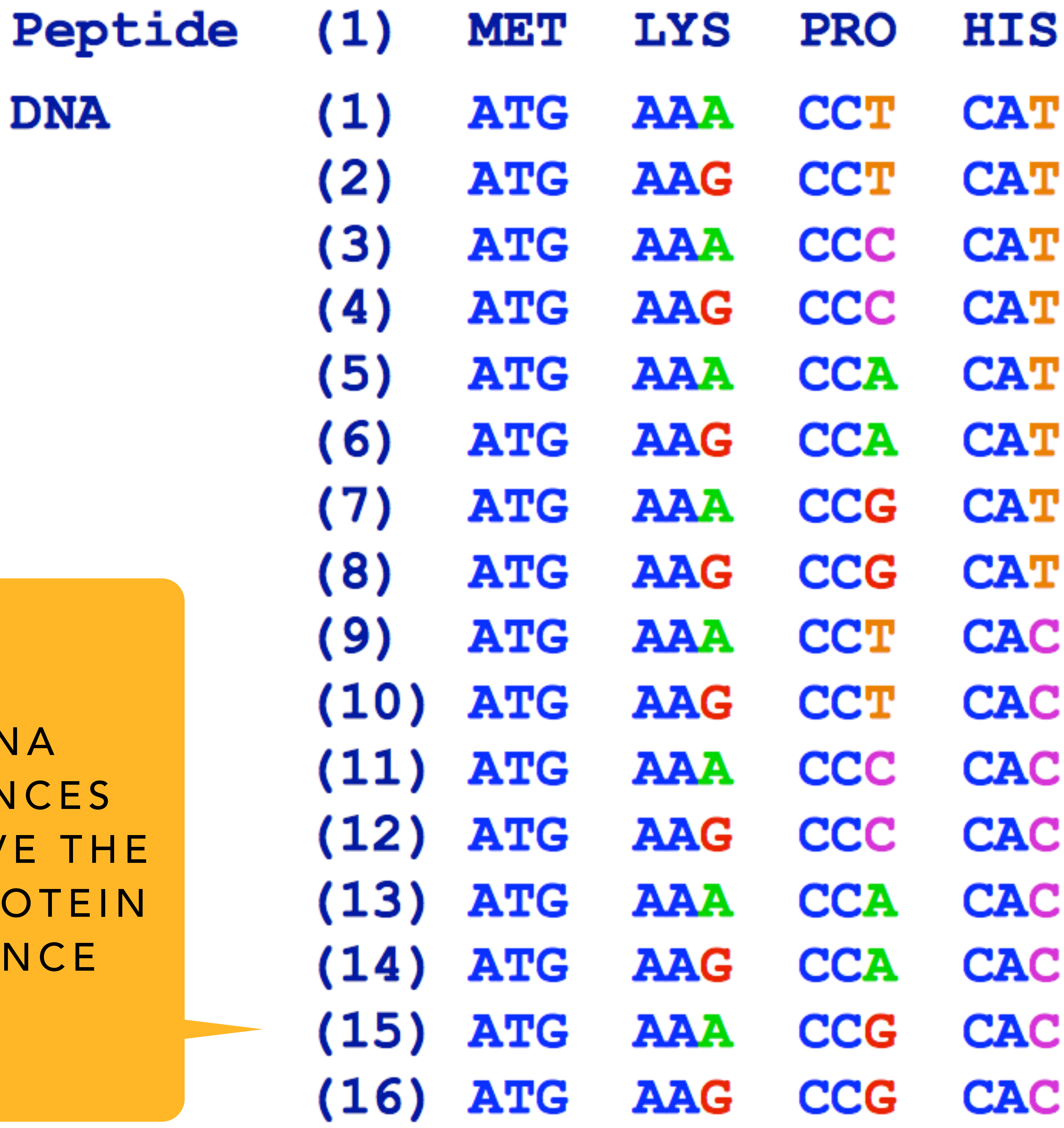
HEURISTIC SEQUENCE ALIGNMENT

- Global or Local?
  - Both local and global alignment methods may be applied to database searching
  - Local alignment methods are more useful since they do not make the assumption that the query protein and database sequence are of similar length

DNA vs Protein Searches

DNA VS PROTEIN SEARCHES

- Protein similarity infers homology
  - Conserved alphabet
- DNA similarity has less sensitivity
  - Reduced alphabet



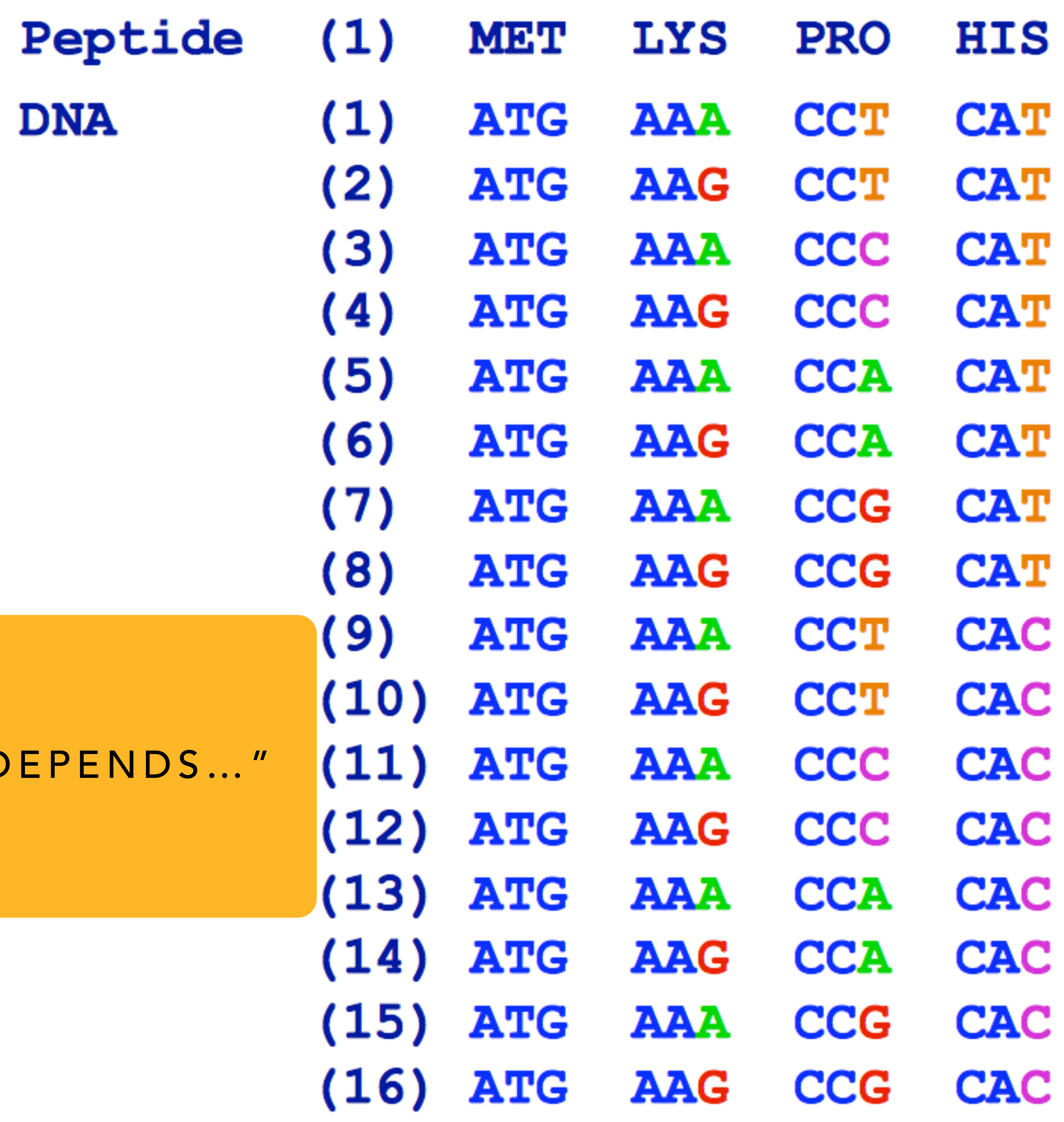
Peptide	(1)	MET	LYS	PRO	HIS
DNA	(1)	ATG	AAA	CCT	CAT
	(2)	ATG	AAG	CCT	CAT
	(3)	ATG	AAA	CCC	CAT
	(4)	ATG	AAG	CCC	CAT
	(5)	ATG	AAA	CCA	CAT
	(6)	ATG	AAG	CCA	CAT
	(7)	ATG	AAA	CCG	CAT
	(8)	ATG	AAG	CCG	CAT
	(9)	ATG	AAA	CCT	CAC
	(10)	ATG	AAG	CCT	CAC
	(11)	ATG	AAA	CCC	CAC
	(12)	ATG	AAG	CCC	CAC
	(13)	ATG	AAA	CCA	CAC
	(14)	ATG	AAG	CCA	CAC
	(15)	ATG	AAA	CCG	CAC
	(16)	ATG	AAG	CCG	CAC

16 DNA SEQUENCES THAT GIVE THE SAME PROTEIN SEQUENCE

DNA VS PROTEIN SEARCHES

- Translate DNA to protein before search?
  - Inherent information loss from degenerate codons
  - Different DNA sequences can encode same protein

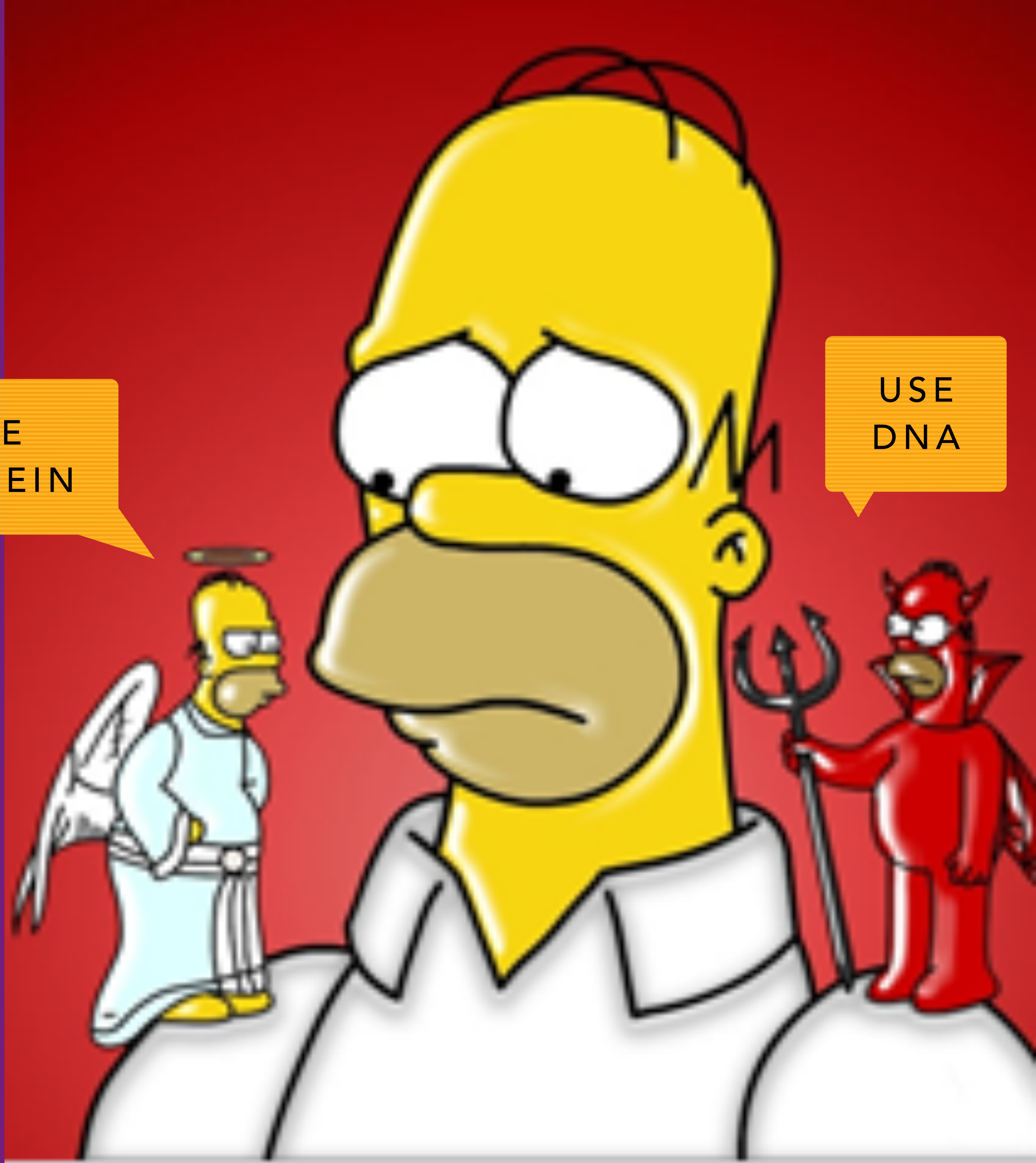
“IT DEPENDS…”



Peptide	(1)	MET	LYS	PRO	HIS
DNA	(1)	ATG	AAA	CCT	CAT
	(2)	ATG	AAG	CCT	CAT
	(3)	ATG	AAA	CCC	CAT
	(4)	ATG	AAG	CCC	CAT
	(5)	ATG	AAA	CCA	CAT
	(6)	ATG	AAG	CCA	CAT
	(7)	ATG	AAA	CCG	CAT
	(8)	ATG	AAG	CCG	CAT
	(9)	ATG	AAA	CCT	CAC
	(10)	ATG	AAG	CCT	CAC
	(11)	ATG	AAA	CCC	CAC
	(12)	ATG	AAG	CCC	CAC
	(13)	ATG	AAA	CCA	CAC
	(14)	ATG	AAG	CCA	CAC
	(15)	ATG	AAA	CCG	CAC
	(16)	ATG	AAG	CCG	CAC

DNA VS PROTEIN SEARCHES

- If given the option, use protein sequences for database similarity searches when possible



USE PROTEIN

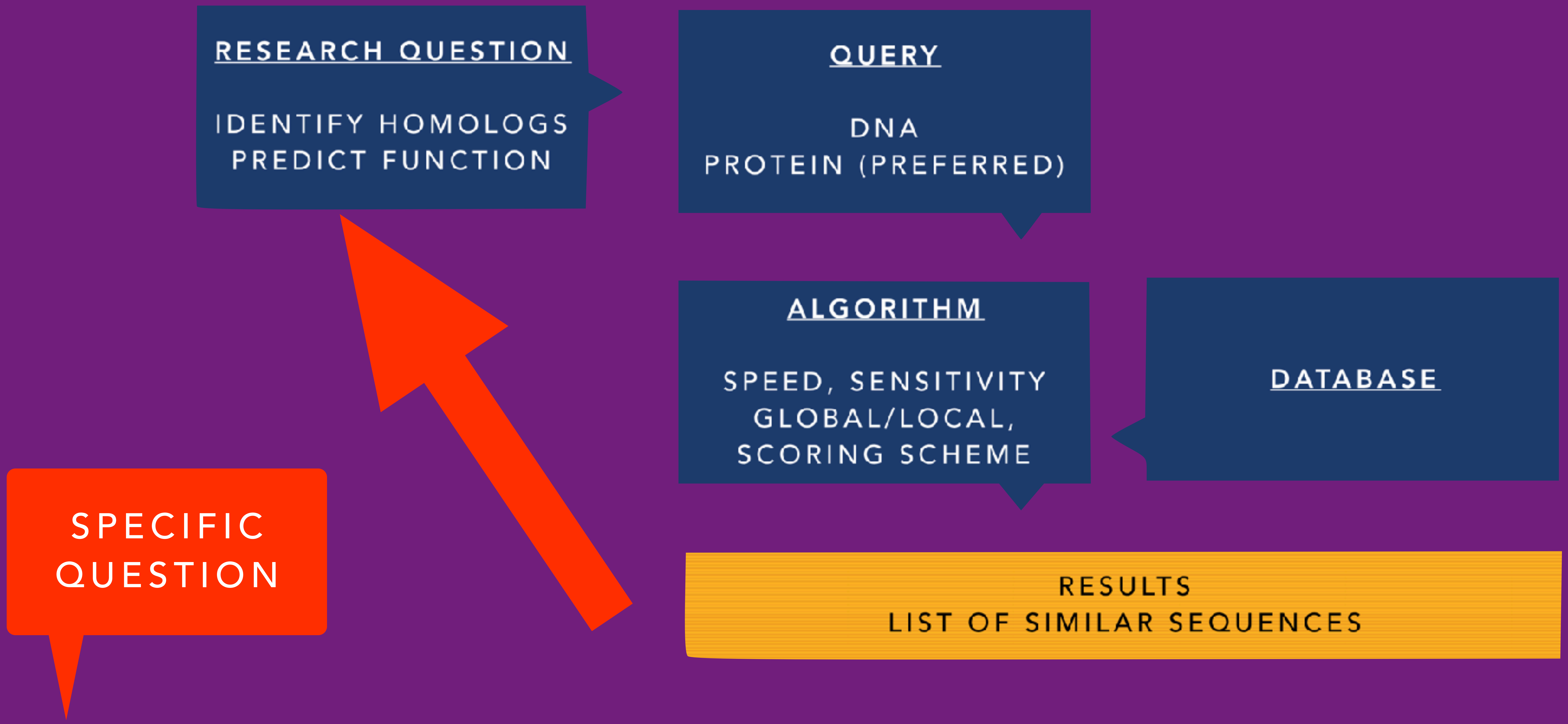
USE DNA

DNA VS PROTEIN SEARCHES

- Why?
  - DNA will have more random matches (bad matches; skew statistics)
  - DNA databases are larger and grow faster
  - DNA uses identity matrices; protein uses scoring matrices (more sensitivity)
  - Protein sequences diverge less than DNA encoding them

Evaluating a Database Search

EVALUATING A DATABASE SEARCH



RESEARCH QUESTION

IDENTIFY HOMOLOGS
PREDICT FUNCTION

QUERY

DNA
PROTEIN (PREFERRED)

ALGORITHM

SPEED, SENSITIVITY
GLOBAL/LOCAL,
SCORING SCHEME

DATABASE

RESULTS
LIST OF SIMILAR SEQUENCES

SPECIFIC QUESTION

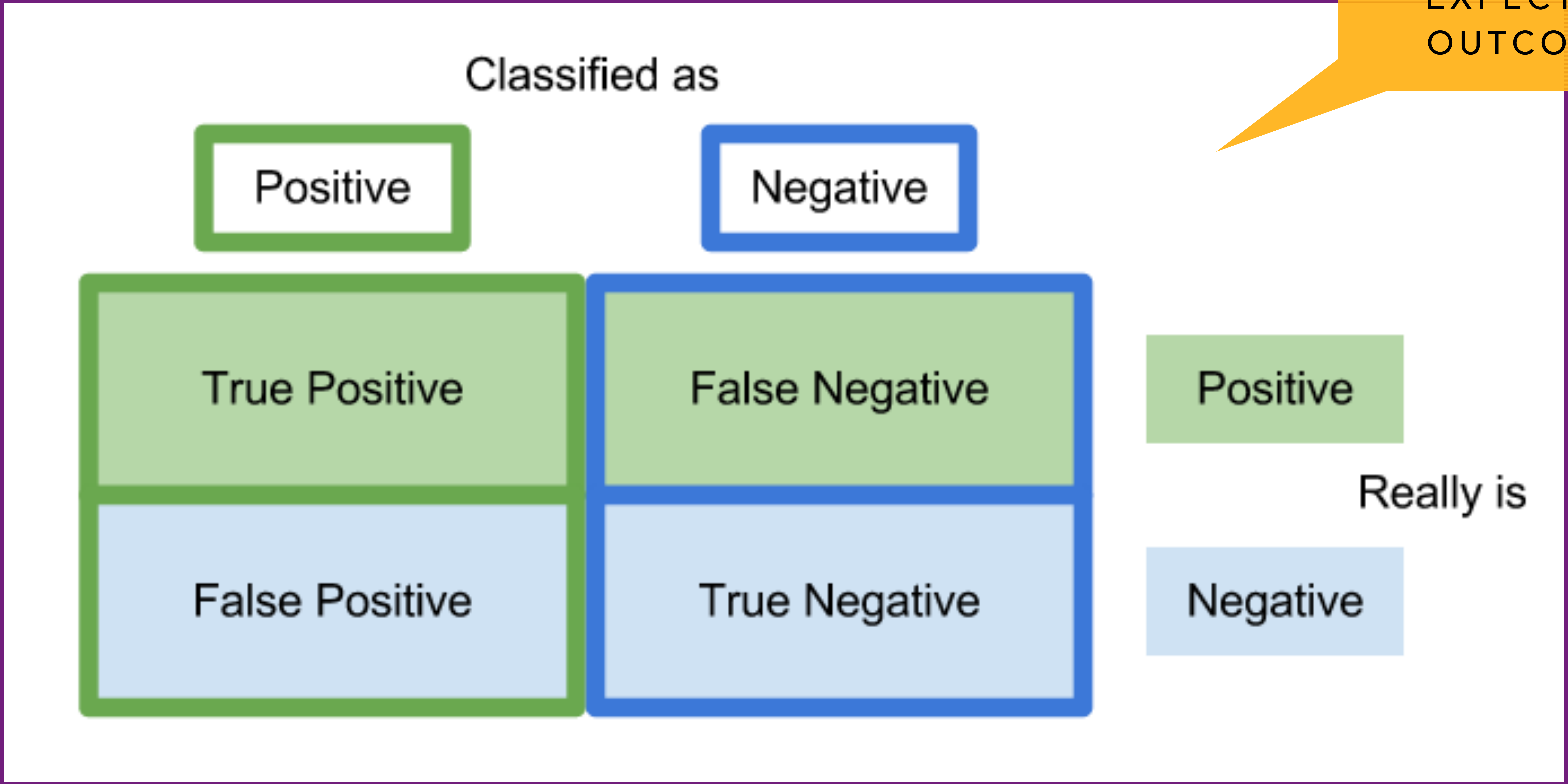
- How can we evaluate a database search?

EVALUATING A DATABASE SEARCH

- If we were to develop a new algorithm, how could we compare it to existing algorithms?

EVALUATING A DATABASE SEARCH

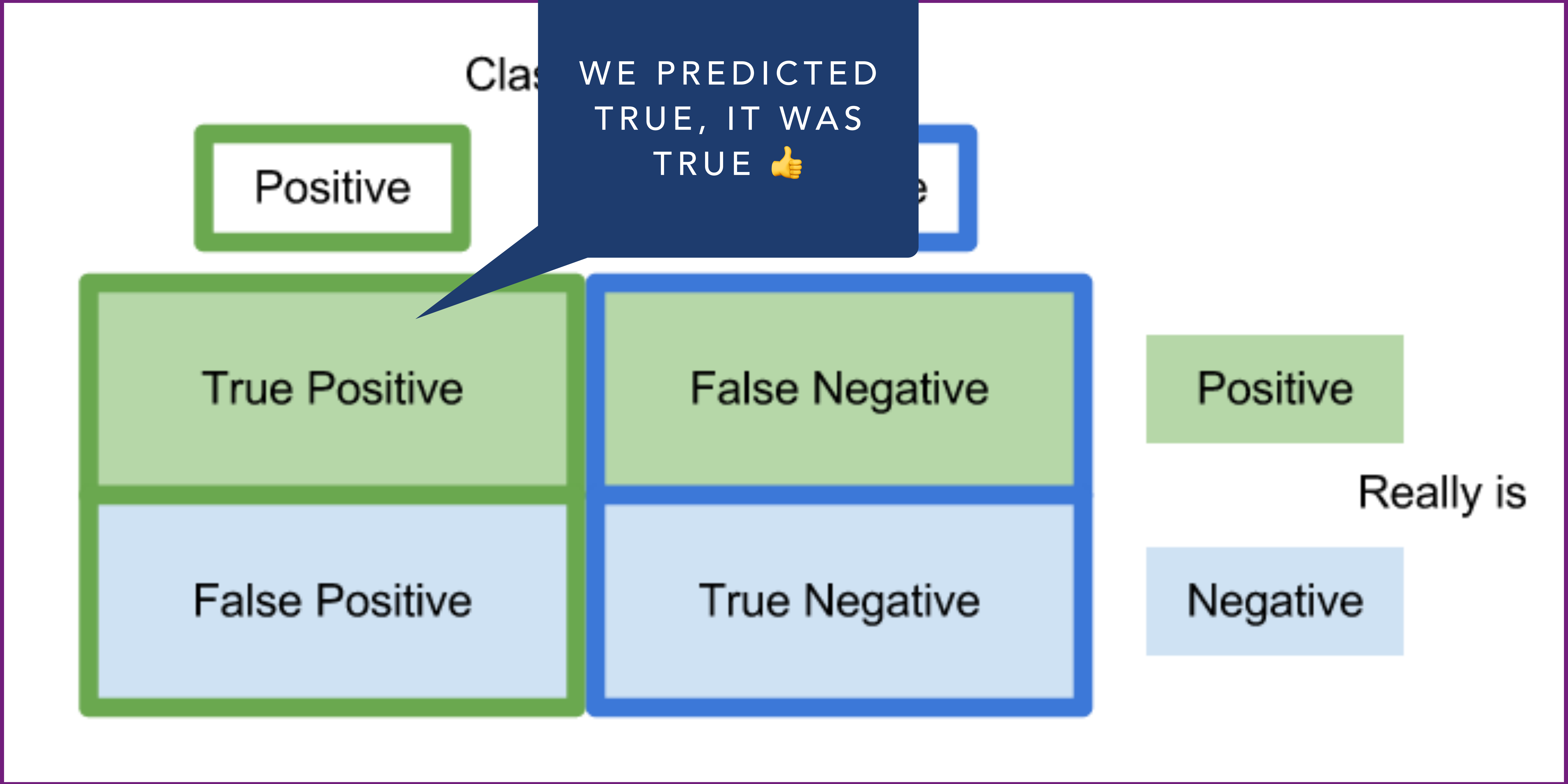
WHEN WE LOOK AT A RESULT COMPARED TO EXPECTED OUTCOME



Classified as

Positive	Negative	Really is
True Positive	False Negative	Positive
False Positive	True Negative	Negative

EVALUATING A DATABASE SEARCH



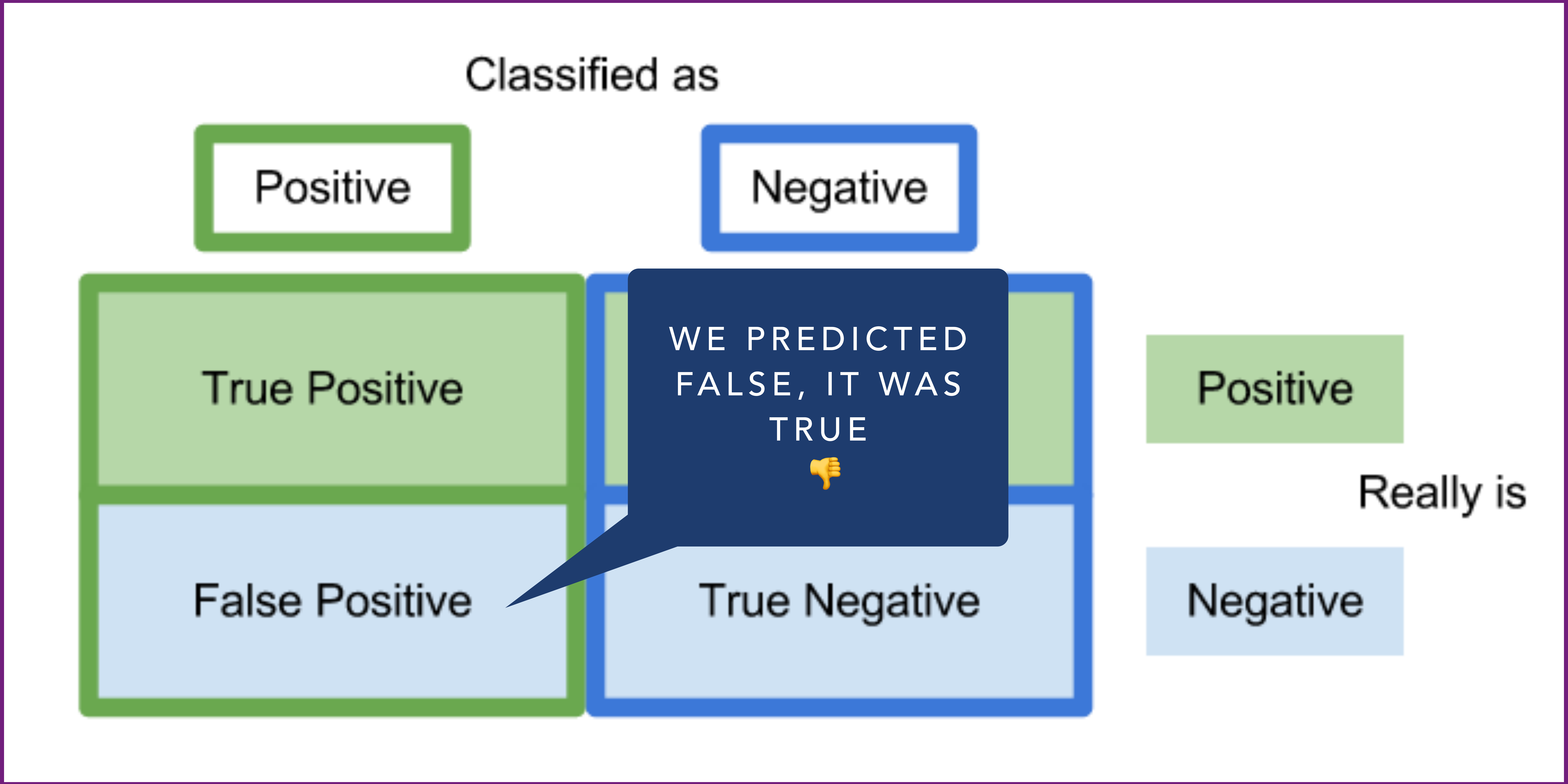
Clas

Positive	e	
True Positive	False Negative	Positive
False Positive	True Negative	Negative

Really is

WE PREDICTED TRUE, IT WAS TRUE 👍

EVALUATING A DATABASE SEARCH

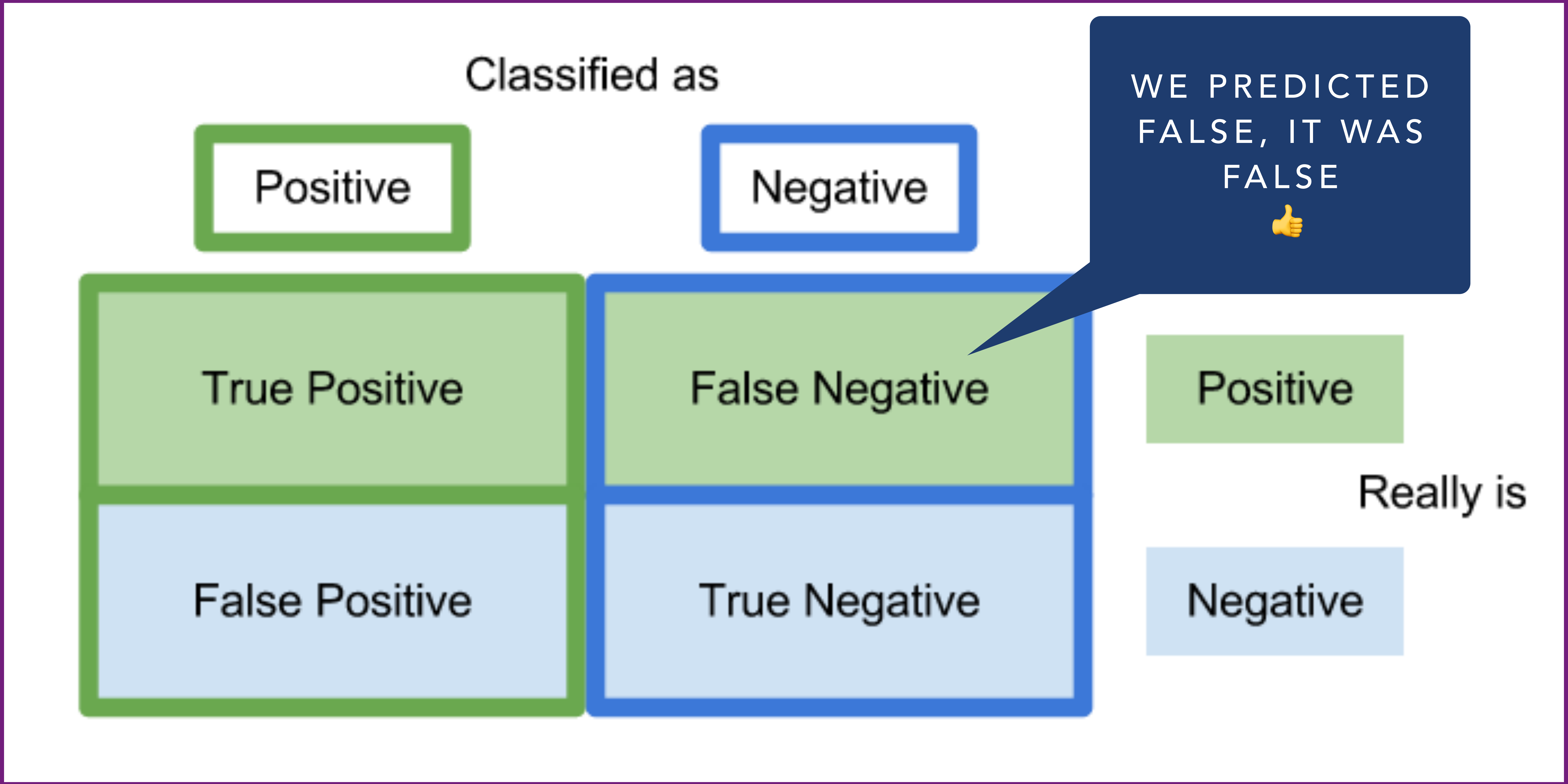


Classified as

Positive	Negative	Really is
True Positive		Positive
False Positive	True Negative	Negative

WE PREDICTED FALSE, IT WAS TRUE 👎

EVALUATING A DATABASE SEARCH



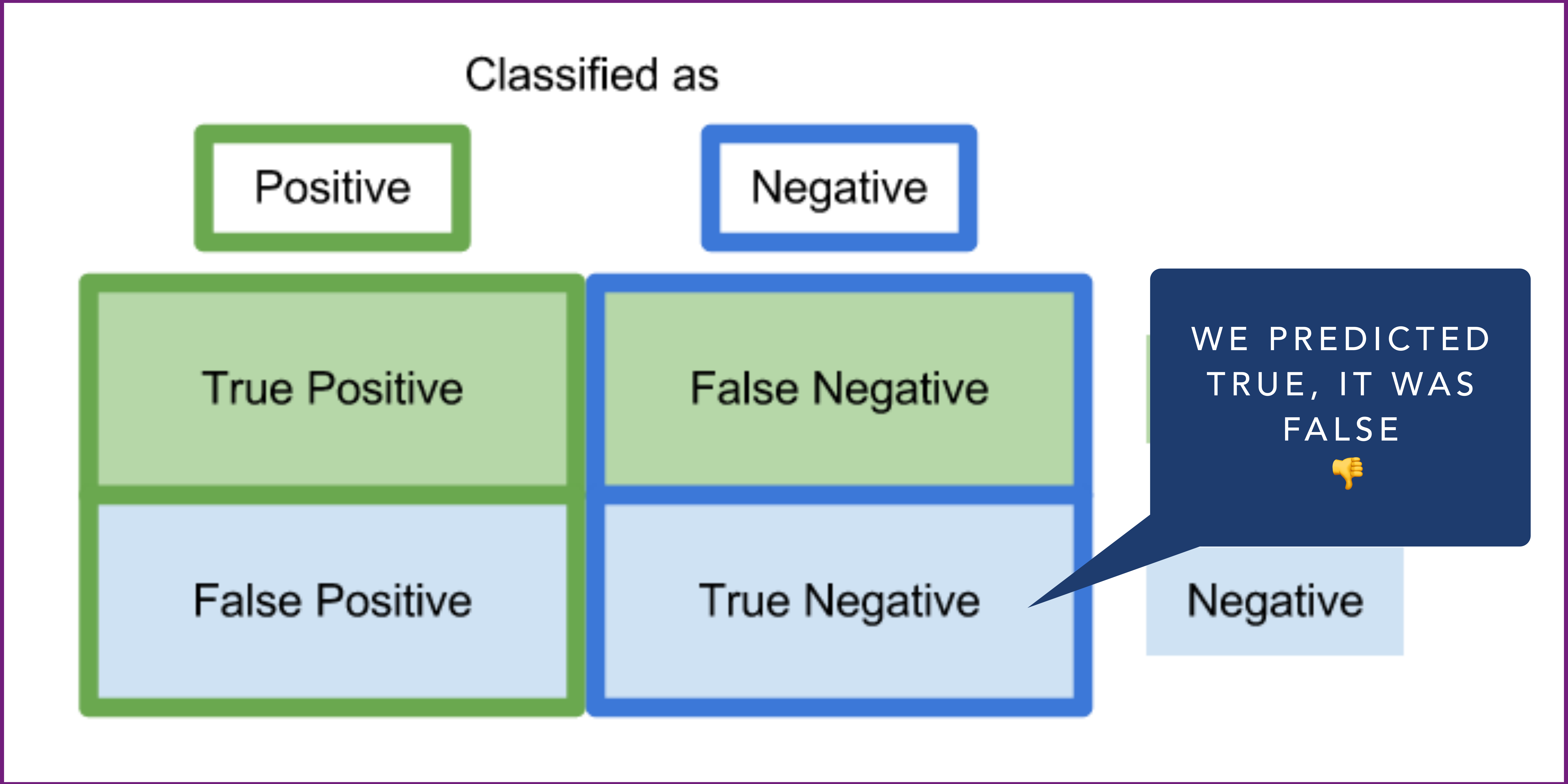
Classified as

	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Really is

WE PREDICTED FALSE, IT WAS FALSE
👍

EVALUATING A DATABASE SEARCH



Classified as

Positive	Negative	
True Positive	False Negative	
False Positive	True Negative	Negative

WE PREDICTED TRUE, IT WAS FALSE 👎

EVALUATING A DATABASE SEARCH

- Example database of food
- Query for “fruit”
- Results are sorted by score
  - > 90 predicted fruit
  - < 90 predicted not a fruit

THRESHOLD IS HEURISTIC

```
> apple     | 100
> orange    |  99
> banana    |  93
> pumpkin   |  93
> grapes    |  92
> eggplant  |  91
> kiwi      |  88
> lettuce   |  85
> tomato    |  79
> onion     |  77
```

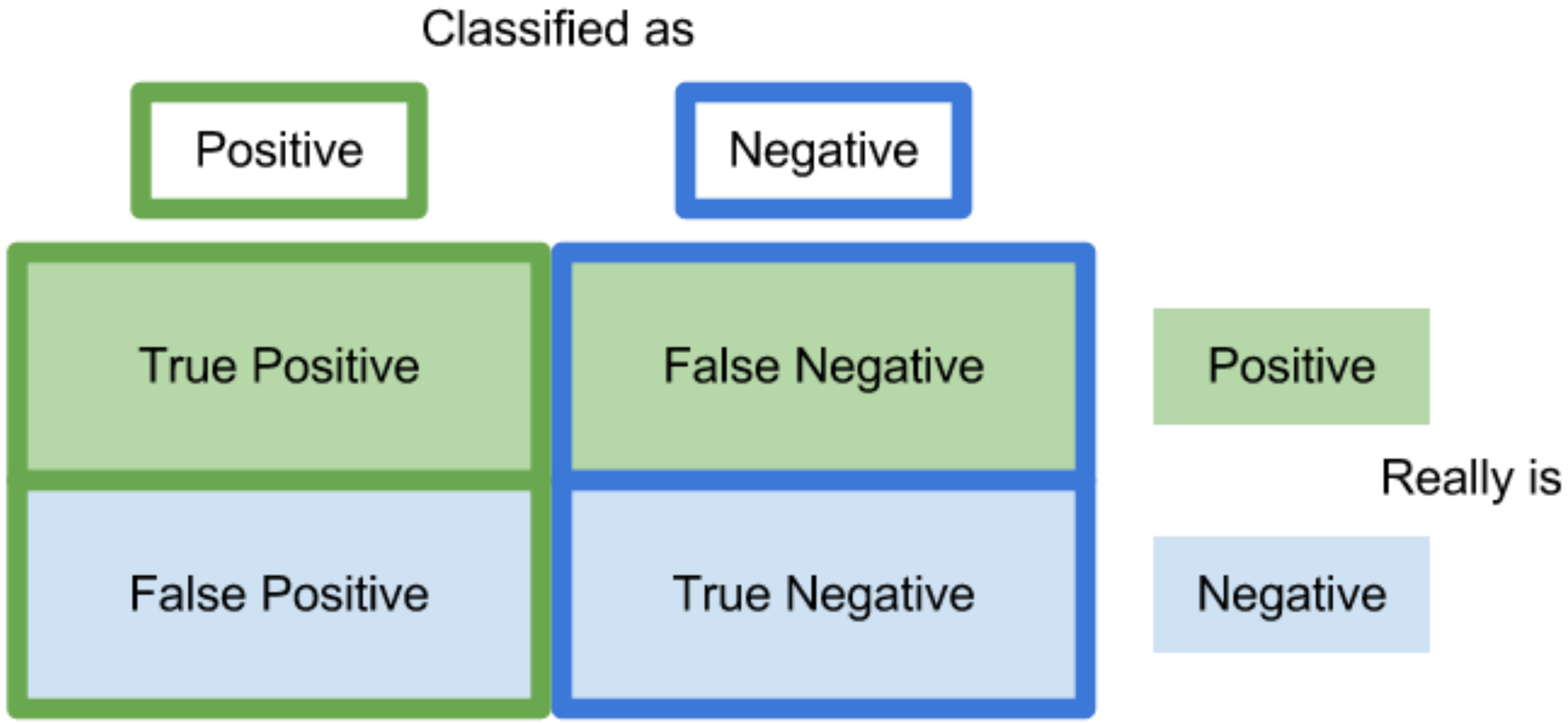
EVALUATING A DATABASE SEARCH

- True Positive - Predicted correctly
- False Positive - Predicted wrong
- True Negative - Correctly predicted not a member
- False Negative - Missed it

```
                       TP
                       FP
                       TP
> apple     | 100 FP
> orange    | 99  FN
> banana    | 93  TN
> pumpkin   | 93  FN
> grapes    | 92  TN
> eggplant  | 91
> kiwi      | 88
> lettuce   | 85
> tomato    | 79
> onion     | 77
```

EVALUATING A DATABASE SEARCH

- Sensitivity
  - Ability to correctly classify as homologous
  - Sensitivity = TP/TP+FN
- Specificity
  - Ability to correctly classify as non-homologous
  - Specificity = TN/(TN+FP)



Classified as

Really is	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

EVALUATING A DATABASE SEARCH

TRUE

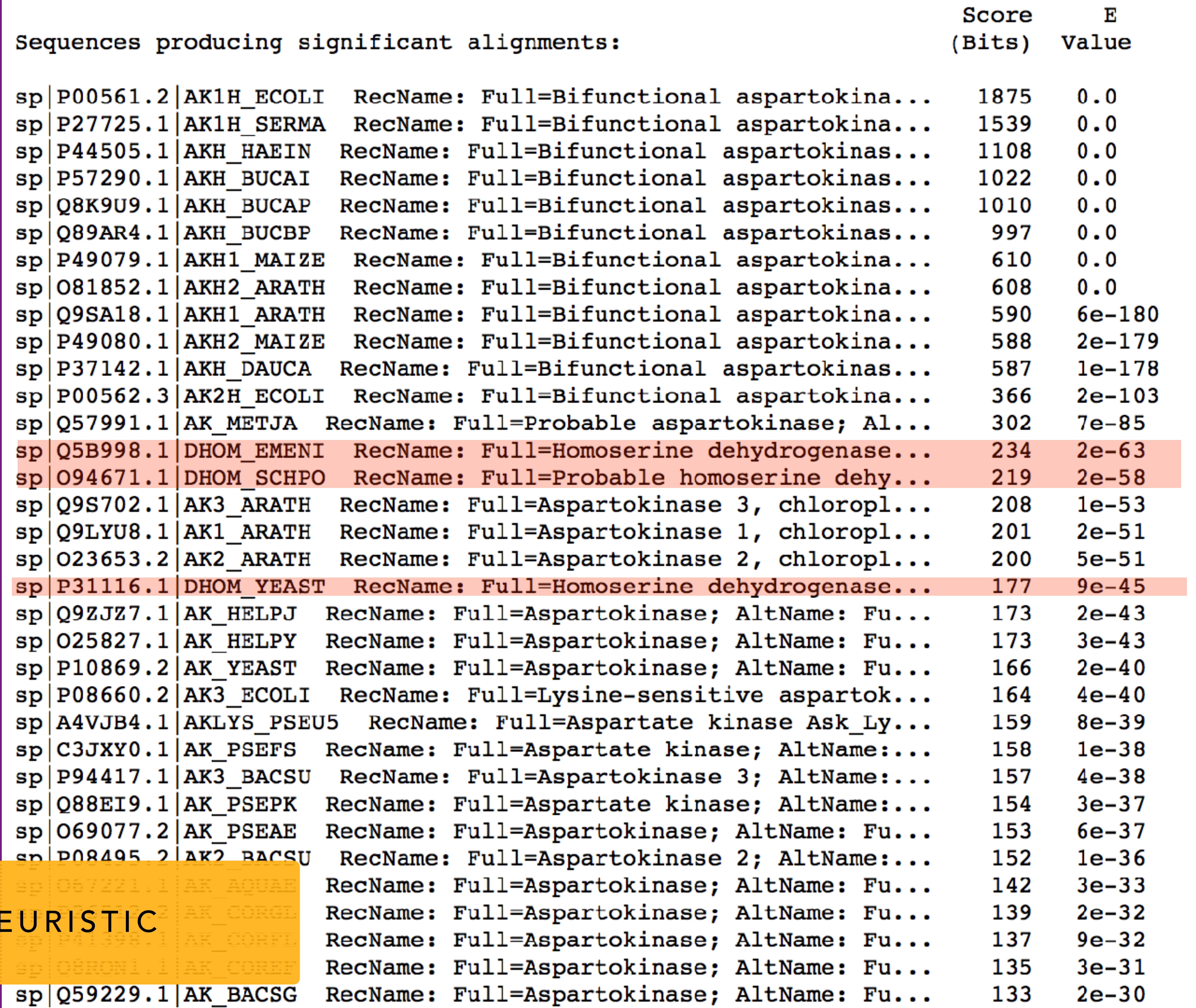
- Sensitivity = TP/TP+FN
  - 4/6 = 66%
- Specificity = TN/(TN+FP)
  - 2/4 = 50%

ACCURATE
PREDICT
NOT TRUE

```
                            TP
                            FP
                            TP
> apple       |  100  FP
> orange      |  99   FN
> banana      |  93   TN
> pumpkin     |  93   FN
> grapes      |  92   TN
> eggplant    |  91
> kiwi        |  88
> lettuce     |  85
> tomato      |  79
> onion       |  77
```

DATABASE SEARCHING

- BLAST results for query of "aspartokinase
- Determine "true" by name
- As precise as you want to be

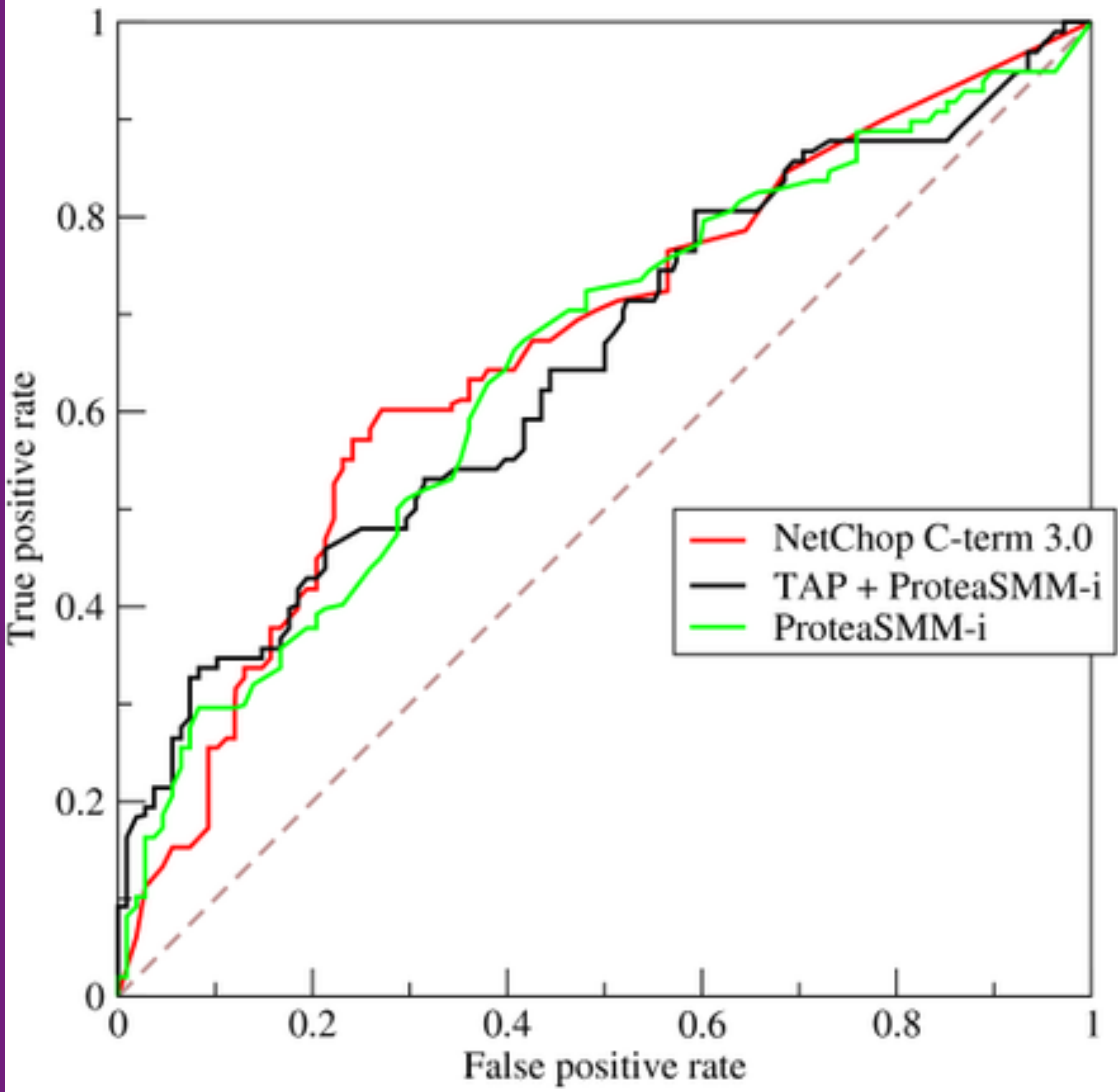


```
                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

sp|P00561.2|AK1H_ECOLI  RecName: Full=Bifunctional aspartokina...   1875    0.0
sp|P27725.1|AK1H_SERMA  RecName: Full=Bifunctional aspartokina...   1539    0.0
sp|P44505.1|AKH_HAEIN  RecName: Full=Bifunctional aspartokinas...   1108    0.0
sp|P57290.1|AKH_BUCAI  RecName: Full=Bifunctional aspartokinas...   1022    0.0
sp|Q8K9U9.1|AKH_BUCAP  RecName: Full=Bifunctional aspartokinas...   1010    0.0
sp|Q89AR4.1|AKH_BUCBP  RecName: Full=Bifunctional aspartokinas...    997    0.0
sp|P49079.1|AKH1_MAIZE  RecName: Full=Bifunctional aspartokina...    610    0.0
sp|O81852.1|AKH2_ARATH  RecName: Full=Bifunctional aspartokina...    608    0.0
sp|Q9SA18.1|AKH1_ARATH  RecName: Full=Bifunctional aspartokina...    590    6e-180
sp|P49080.1|AKH2_MAIZE  RecName: Full=Bifunctional aspartokina...    588    2e-179
sp|P37142.1|AKH_DAUCA  RecName: Full=Bifunctional aspartokinas...    587    1e-178
sp|P00562.3|AK2H_ECOLI  RecName: Full=Bifunctional aspartokina...    366    2e-103
sp|Q57991.1|AK_METJA  RecName: Full=Probable aspartokinase; Al...    302    7e-85
sp|Q5B998.1|DHOM_EMENI  RecName: Full=Homoserine dehydrogenase...    234    2e-63
sp|O94671.1|DHOM_SCHPO  RecName: Full=Probable homoserine dehy...    219    2e-58
sp|Q9S702.1|AK3_ARATH  RecName: Full=Aspartokinase 3, chloropl...    208    1e-53
sp|Q9LYU8.1|AK1_ARATH  RecName: Full=Aspartokinase 1, chloropl...    201    2e-51
sp|O23653.2|AK2_ARATH  RecName: Full=Aspartokinase 2, chloropl...    200    5e-51
sp|P31116.1|DHOM_YEAST  RecName: Full=Homoserine dehydrogenase...    177    9e-45
sp|Q9ZJZ7.1|AK_HELPJ  RecName: Full=Aspartokinase; AltName: Fu...    173    2e-43
sp|O25827.1|AK_HELPY  RecName: Full=Aspartokinase; AltName: Fu...    173    3e-43
sp|P10869.2|AK_YEAST  RecName: Full=Aspartokinase; AltName: Fu...    166    2e-40
sp|P08660.2|AK3_ECOLI  RecName: Full=Lysine-sensitive aspartok...    164    4e-40
sp|A4VJB4.1|AKLYS_PSEU5  RecName: Full=Aspartate kinase Ask_Ly...    159    8e-39
sp|C3JXY0.1|AK_PSEFS  RecName: Full=Aspartate kinase; AltName:...    158    1e-38
sp|P94417.1|AK3_BACSU  RecName: Full=Aspartokinase 3; AltName:...    157    4e-38
sp|Q88EI9.1|AK_PSEPK  RecName: Full=Aspartate kinase; AltName:...    154    3e-37
sp|O69077.2|AK_PSEAE  RecName: Full=Aspartokinase; AltName: Fu...    153    6e-37
sp|P08495.2|AK2_BACSU  RecName: Full=Aspartokinase 2; AltName:...    152    1e-36
sp|O67221.1|AK_AQUAE  RecName: Full=Aspartokinase; AltName: Fu...    142    3e-33
sp|P26512.2|AK_CORGL  RecName: Full=Aspartokinase; AltName: Fu...    139    2e-32
sp|P41398.1|AK_CORFL  RecName: Full=Aspartokinase; AltName: Fu...    137    9e-32
sp|Q8RQN1.1|AK_COREF  RecName: Full=Aspartokinase; AltName: Fu...    135    3e-31
sp|Q59229.1|AK_BACSG  RecName: Full=Aspartokinase; AltName: Fu...    133    2e-30
```

HEURISTIC

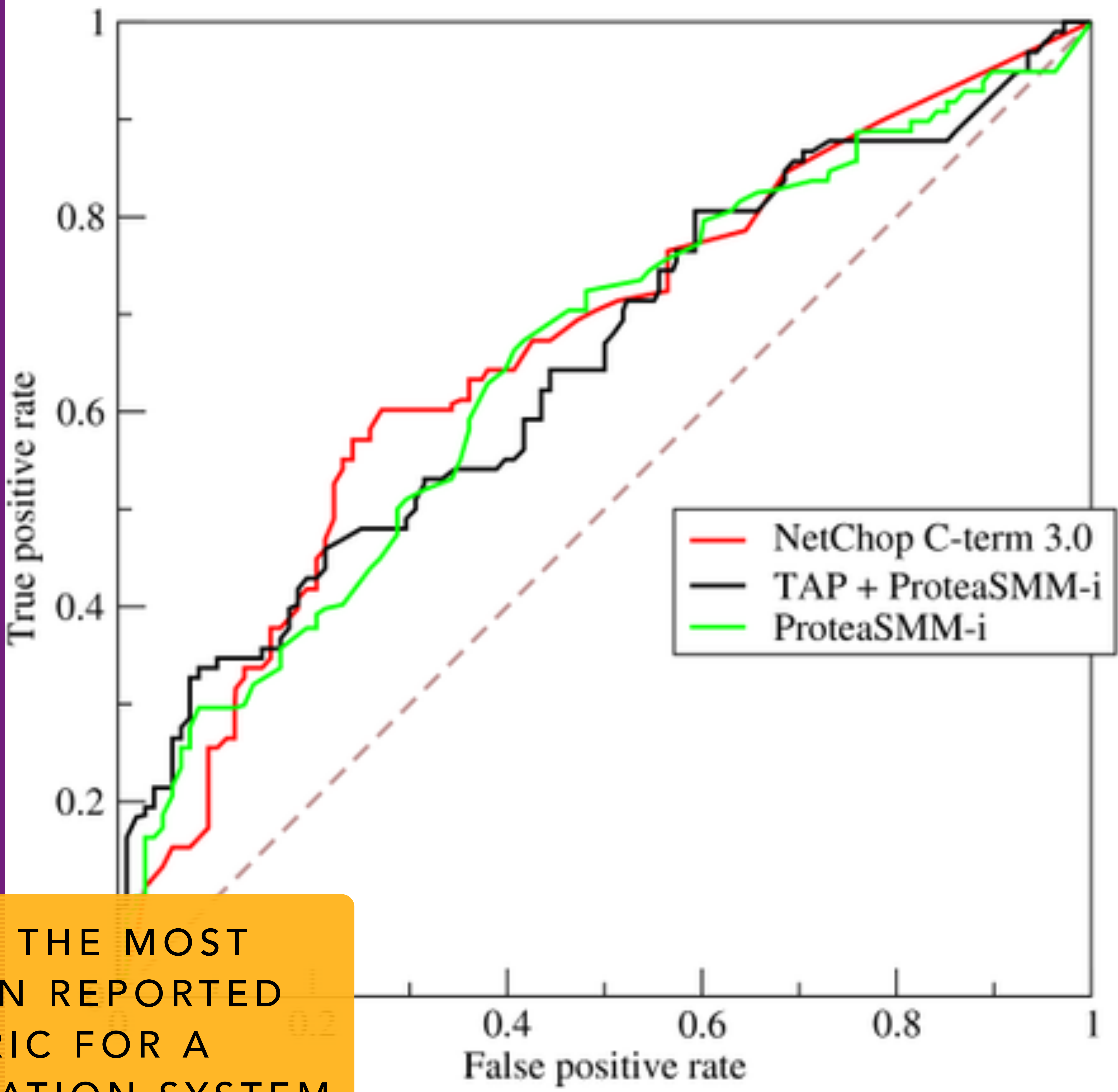
EVALUATING A DATABASE SEARCH

- Receiver operating characteristic (ROC) curve
  - Plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied
  - Sensitivity vs. Specificity
    - TPR/FPR



EVALUATING A DATABASE SEARCH

- Area under curve (AUC)
  - Probability that a randomly choses positive instance over a chosen negative instance
  - Statistic for a classification method
  - http://docs.eyesopen.com/toolkits/cookbook/python/plotting/roc.html



AUC IS THE MOST COMMON REPORTED METRIC FOR A CLASSIFICATION SYSTEM

DATABASE SEARCHING

	Tp	Fp
5	5	0
5-10	5	0
10-15	3	2
15-20	1	4
20-25	3	2
	21	9

OBSERVED TP/FP IN CHUNKS OF 5 TO FIND RATE

```
                                                                   Score     E
Sequences producing significant alignme                           (Bits)  Value

sp|P00561.2|AK1H_ECOLI  RecName: Full=B
sp|P27725.1|AK1H_SERMA  RecName: Full=Bifunctional aspartokina...
sp|P44505.1|AKH_HAEIN  RecName: Full=Bifunctional aspartokinas...
sp|P57290.1|AKH_BUCAI  RecName: Full=Bifunctional aspartokinas...
sp|Q8K9U9.1|AKH_BUCAP  RecName: Full=Bifunctional aspartokinas...
sp|Q89AR4.1|AKH_BUCBP  RecName: Full=Bifunctional aspartokinas...
sp|P49079.1|AKH1_MAIZE  RecName: Full=Bifunctional aspartokina...
sp|O81852.1|AKH2_ARATH  RecName: Full=Bifunctional aspartokina...
sp|Q9SA18.1|AKH1_ARATH  RecName: Full=Bifunctional aspartokina...   590    6e-180
sp|P49080.1|AKH2_MAIZE  RecName: Full=Bifunctional aspartokina...   588    2e-179
sp|P37142.1|AKH_DAUCA  RecName: Full=Bifunctional aspartokinas...   587    1e-178
sp|P00562.3|AK2H_ECOLI  RecName: Full=Bifunctional aspartokina...   366    2e-103
sp|Q57991.1|AK_METJA  RecName: Full=Probable aspartokinase; Al...   302    7e-85
sp|Q5B998.1|DHOM_EMENI  RecName: Full=Homoserine dehydrogenase...   234    2e-63
sp|O94671.1|DHOM_SCHPO  RecName: Full=Probable homoserine dehy...   219    2e-58
sp|Q9S702.1|AK3_ARATH  RecName: Full=Aspartokinase 3, chloropl...   208    1e-53
sp|Q9LYU8.1|AK1_ARATH  RecName: Full=Aspartokinase 1, chloropl...   201    2e-51
```

DATABASE SEARCHING

OBSERVED TP/FP
IN CHUNKS OF 5

	TP	FP
5	5	0
5-10	5	0
10-15	3	2
15-20	1	4
20-25	3	2
	21	9

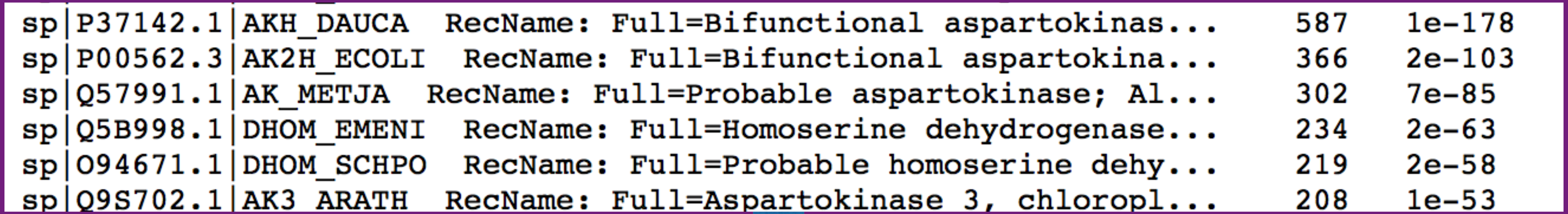
TP/TP_{TOT}

$1-(FP/FP_{TOT})$

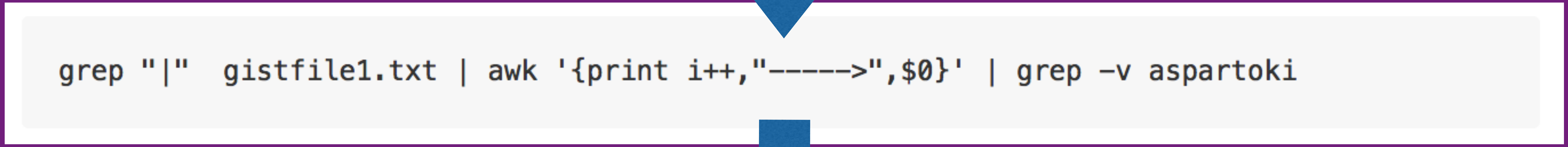
	Sensitivity	Specificity
10	0.24	1
20	0.48	1
30	0.67	0.88
40	0.81	0.67
50	1	0

	TPR	FPR
10	0.24	0
20	0.48	0
30	0.67	0.22
40	0.81	0.33
50	1	1

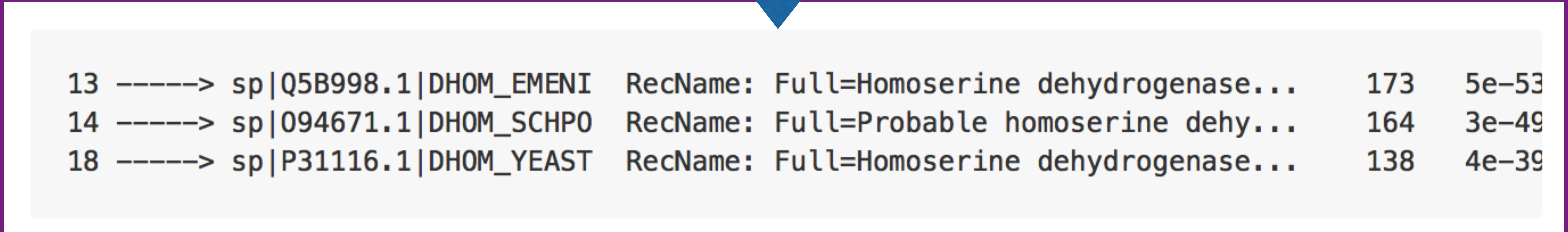
DATABASE SEARCHING



```
sp|P37142.1|AKH_DAUCA  RecName: Full=Bifunctional aspartokinas...    587   1e-178
sp|P00562.3|AK2H_ECOLI  RecName: Full=Bifunctional aspartokina...    366   2e-103
sp|Q57991.1|AK_METJA  RecName: Full=Probable aspartokinase; Al...    302   7e-85
sp|Q5B998.1|DHOM_EMENI  RecName: Full=Homoserine dehydrogenase...    234   2e-63
sp|O94671.1|DHOM_SCHPO  RecName: Full=Probable homoserine dehy...    219   2e-58
sp|Q9S702.1|AK3_ARATH  RecName: Full=Aspartokinase 3, chloropl...    208   1e-53
```

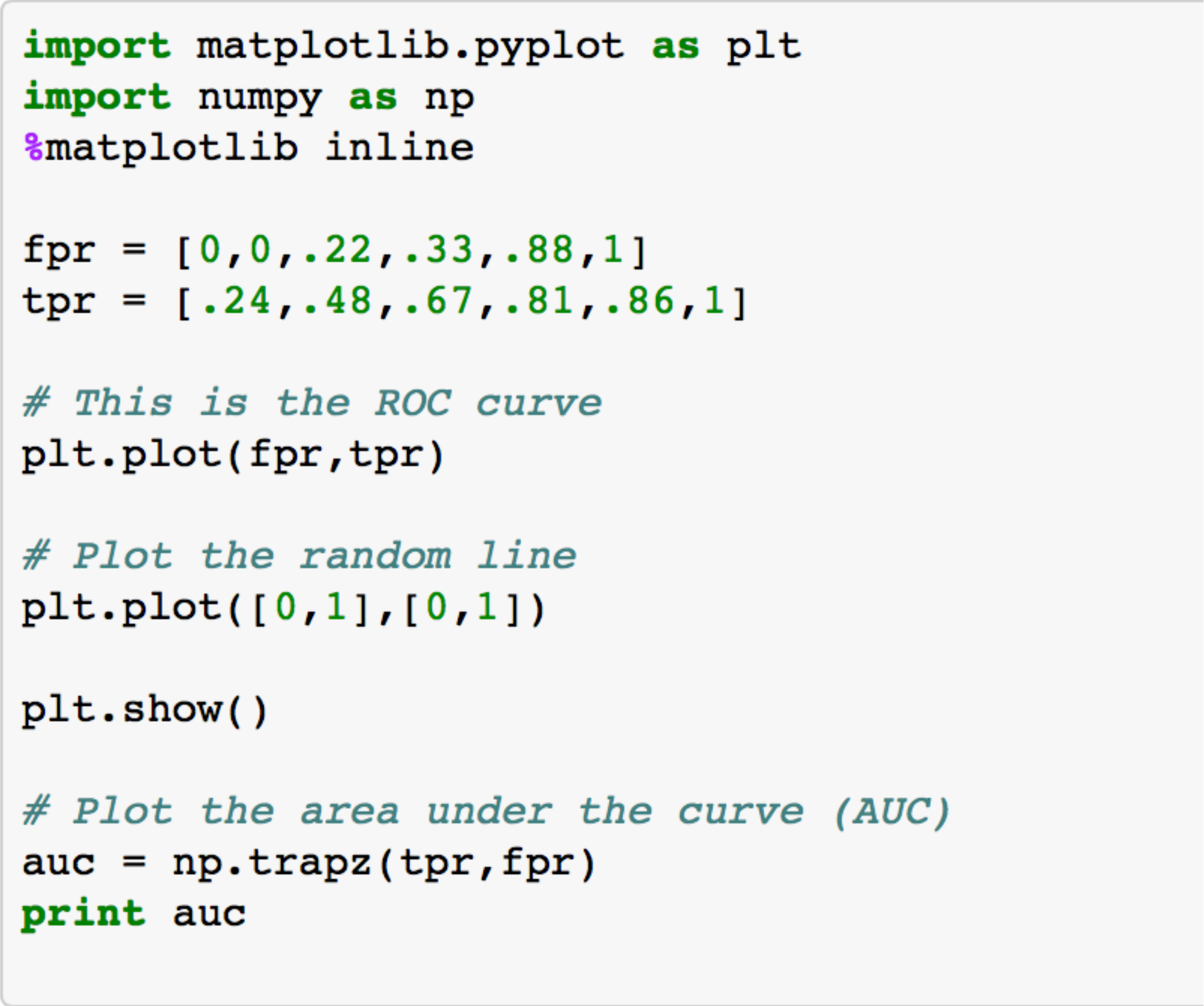


```
grep "|"  gistfile1.txt | awk '{print i++,"----->",$0}' | grep -v aspartoki
```

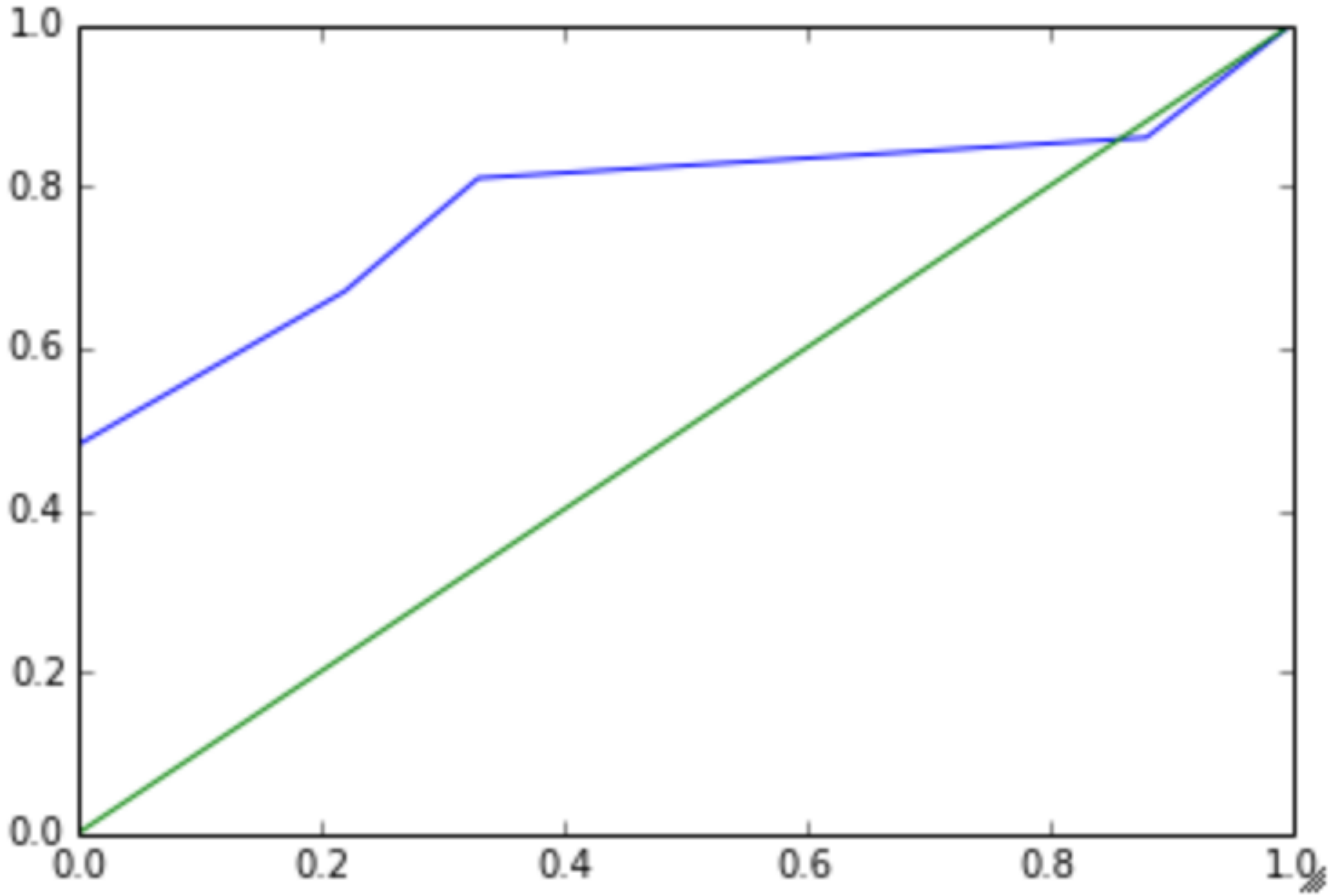


```
13 -----> sp|Q5B998.1|DHOM_EMENI  RecName: Full=Homoserine dehydrogenase...   173   5e-53
14 -----> sp|O94671.1|DHOM_SCHPO  RecName: Full=Probable homoserine dehy...   164   3e-49
18 -----> sp|P31116.1|DHOM_YEAST  RecName: Full=Homoserine dehydrogenase...   138   4e-39
```

DATABASE SEARCHING



```
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline

fpr = [0,0,.22,.33,.88,1]
tpr = [.24,.48,.67,.81,.86,1]

# This is the ROC curve
plt.plot(fpr,tpr)

# Plot the random line
plt.plot([0,1],[0,1])

plt.show()

# Plot the area under the curve (AUC)
auc = np.trapz(tpr,fpr)
print auc
```



```
0.77875
```

BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

MPCS56420
AUTUMN 2018
SESSION 4

THE UNIVERSITY OF
CHICAGO