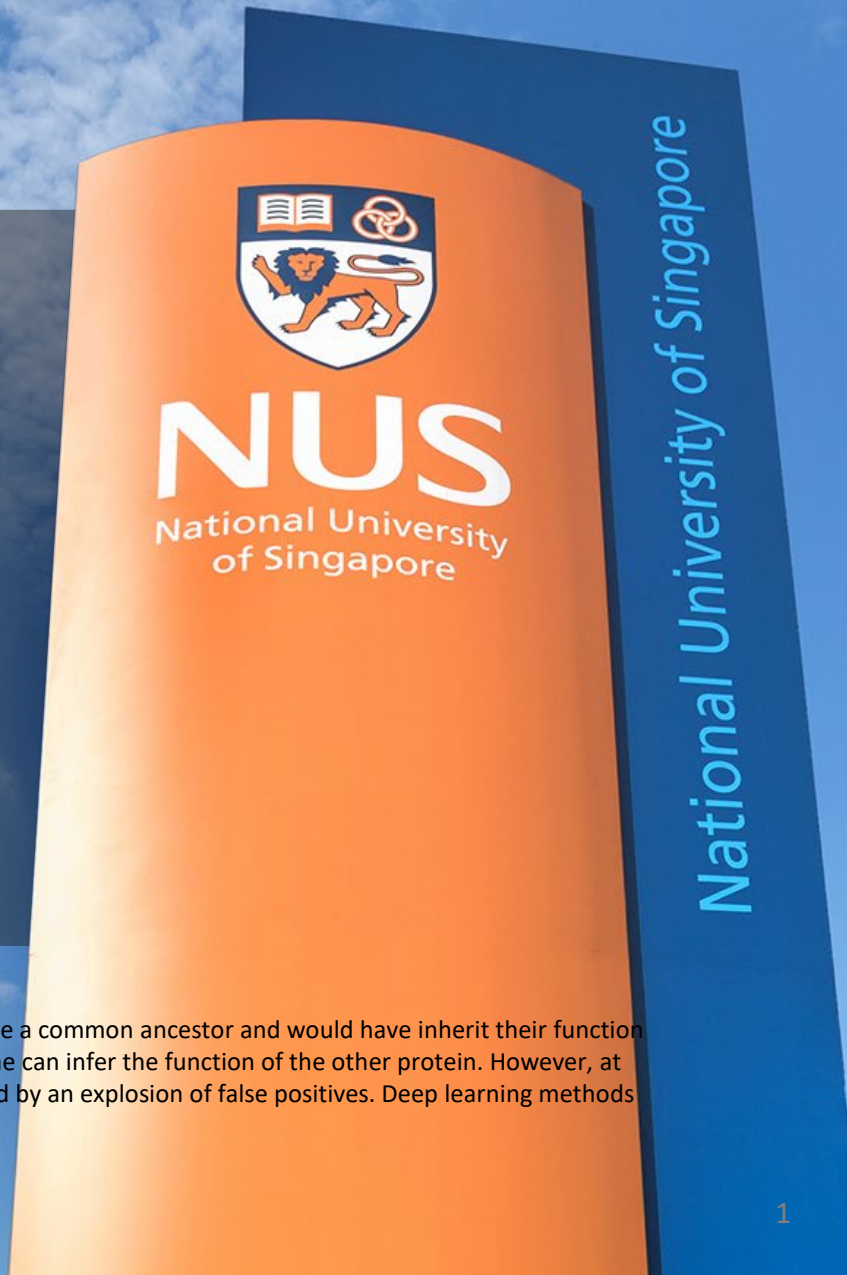
The twilight zone of protein function prediction

Wong Limsoon



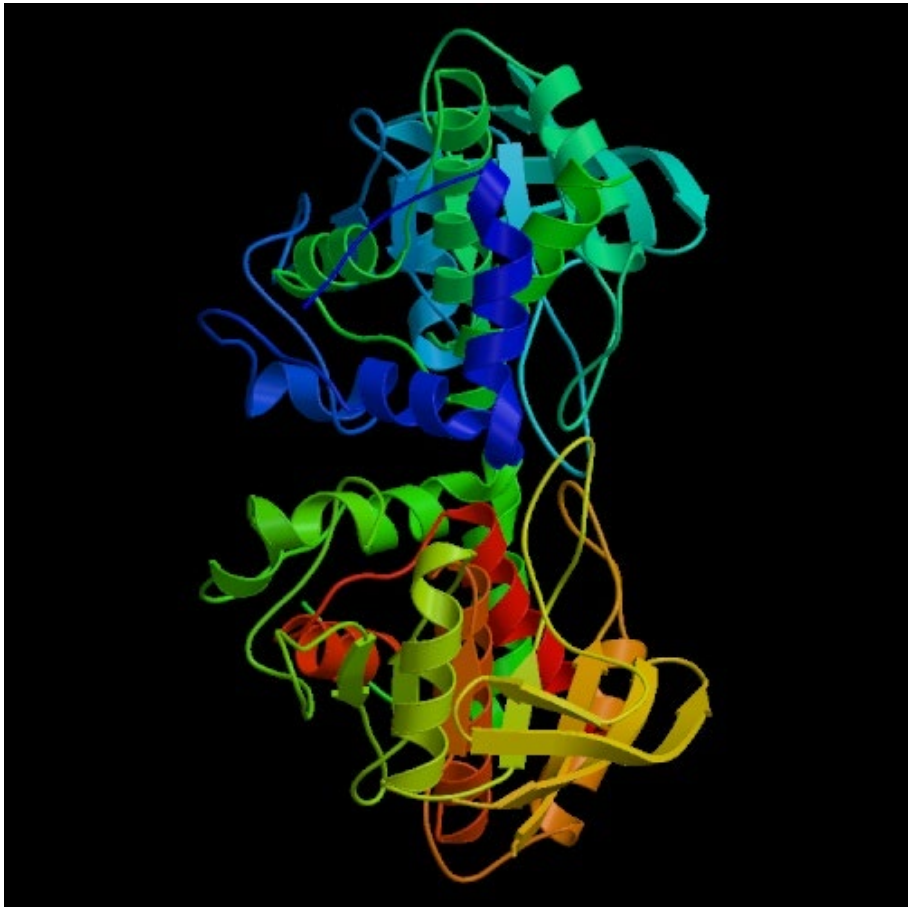
Outline: Generally, if the sequence of two proteins are quite similar, they would have a common ancestor and would have inherit their function from that ancestor. Thus, if one knows the function of one of these two proteins, one can infer the function of the other protein. However, at sequence similarity below 40%, this way of inferring protein function is accompanied by an explosion of false positives. Deep learning methods are proposed as a solution. Do these approaches work?

© Copyright National University of Singapore. All Rights Reserved.

A protein is a ...

A protein is a large complex molecule made up of one or more chains of amino acids

Proteins perform a wide variety of activities in the cell

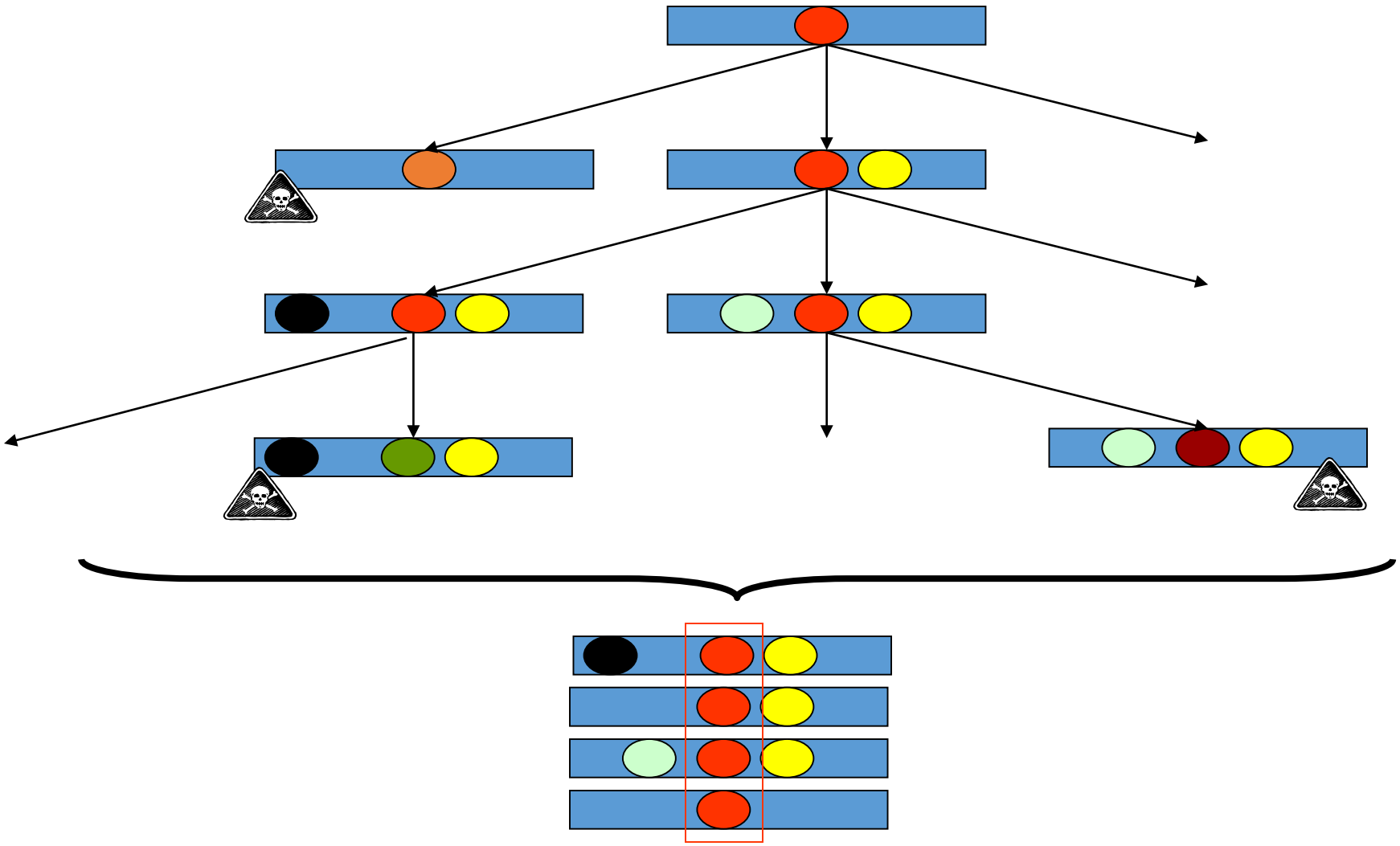


Protein function prediction

SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT

How do we attempt to assign a function to a new protein sequence?

During evolution…



Exercise #1

Let a = AFPHQHRVP

Let b = PQVYNIMKE

Suppose each generation differs from the previous by 1 residue

What is the max difference between the 2nd generation of a

What is the min difference between the 2nd generation of a and b?

Answer

Let a = `AFP HQH RVP`
Let b = `PQV YNI MKE`

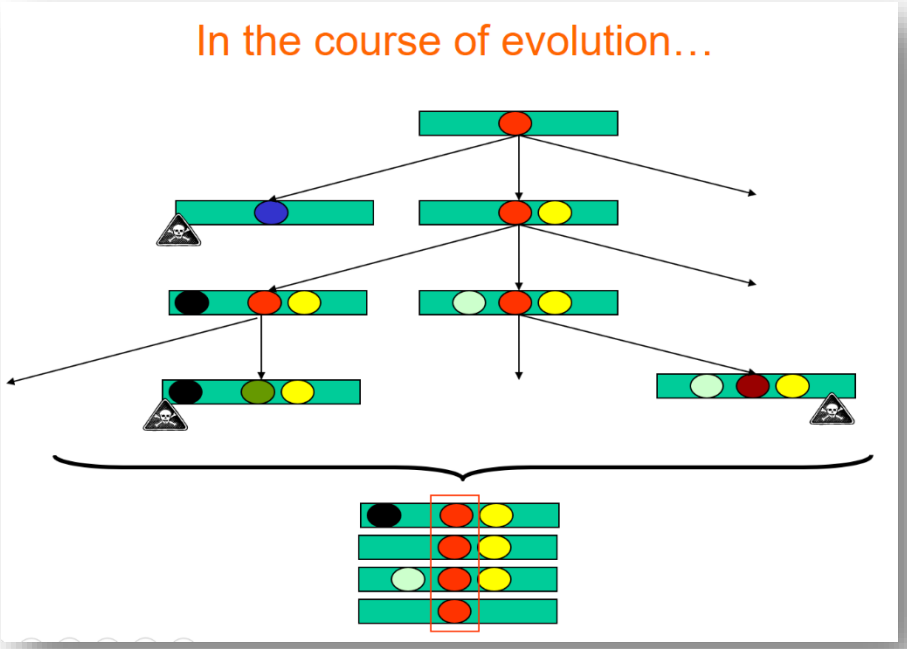
Suppose each generation differs from the previous by 1 residue

What is the max difference between the 2nd generation of a

What is the min difference between the 2nd generation of a and b?

Intentionally left blank

The triumph of logic



Two proteins inheriting their function from a common ancestor have very similar amino acid sequences

Exercise #2

How can we guess the function of a protein?

Abductive reasoning

Intentionally left blank

Guilt by association

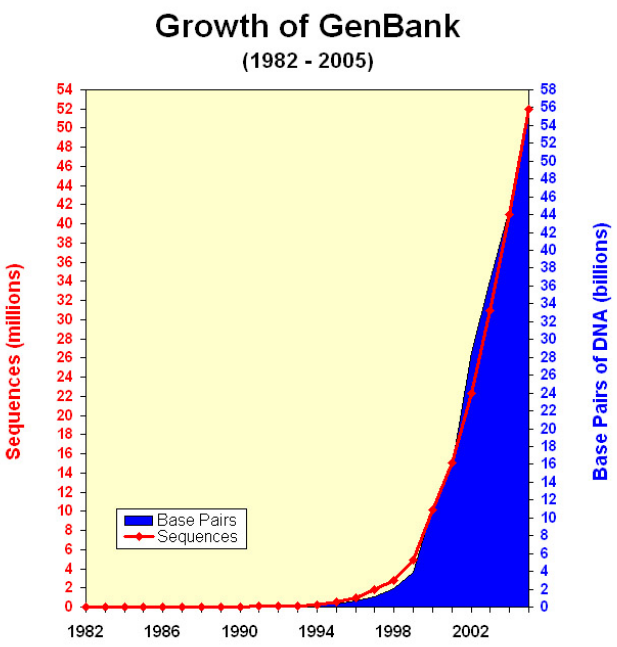
Intentionally left blank

Seq comparison is slow

Time complexity for optimal alignment is $O(n^2)$, where n is seq length

Given size of seq db, use of optimal algo is not practical

⇒BLAST, Pattern Hunter, etc.

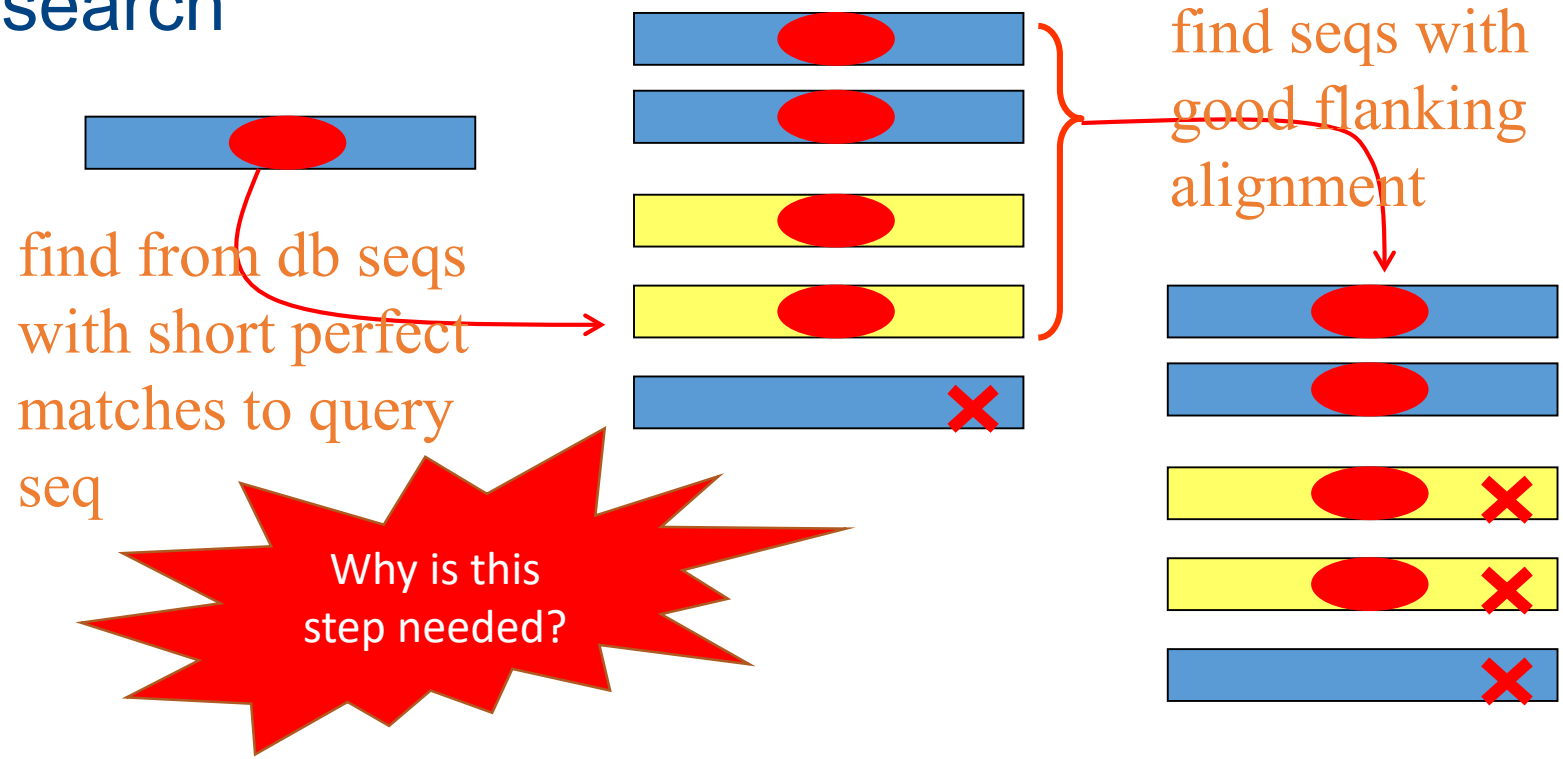


Typical speed-up achieved:
20 min (Smith-Waterman) ➔ 20 sec (BLAST)

BLAST: How it works

Altschul et al., *JMB*, 215:403-410, 1990

BLAST is one of the most popular tool for doing fast "guilt-by-association" sequence homology search



Homologs obtained by BLAST

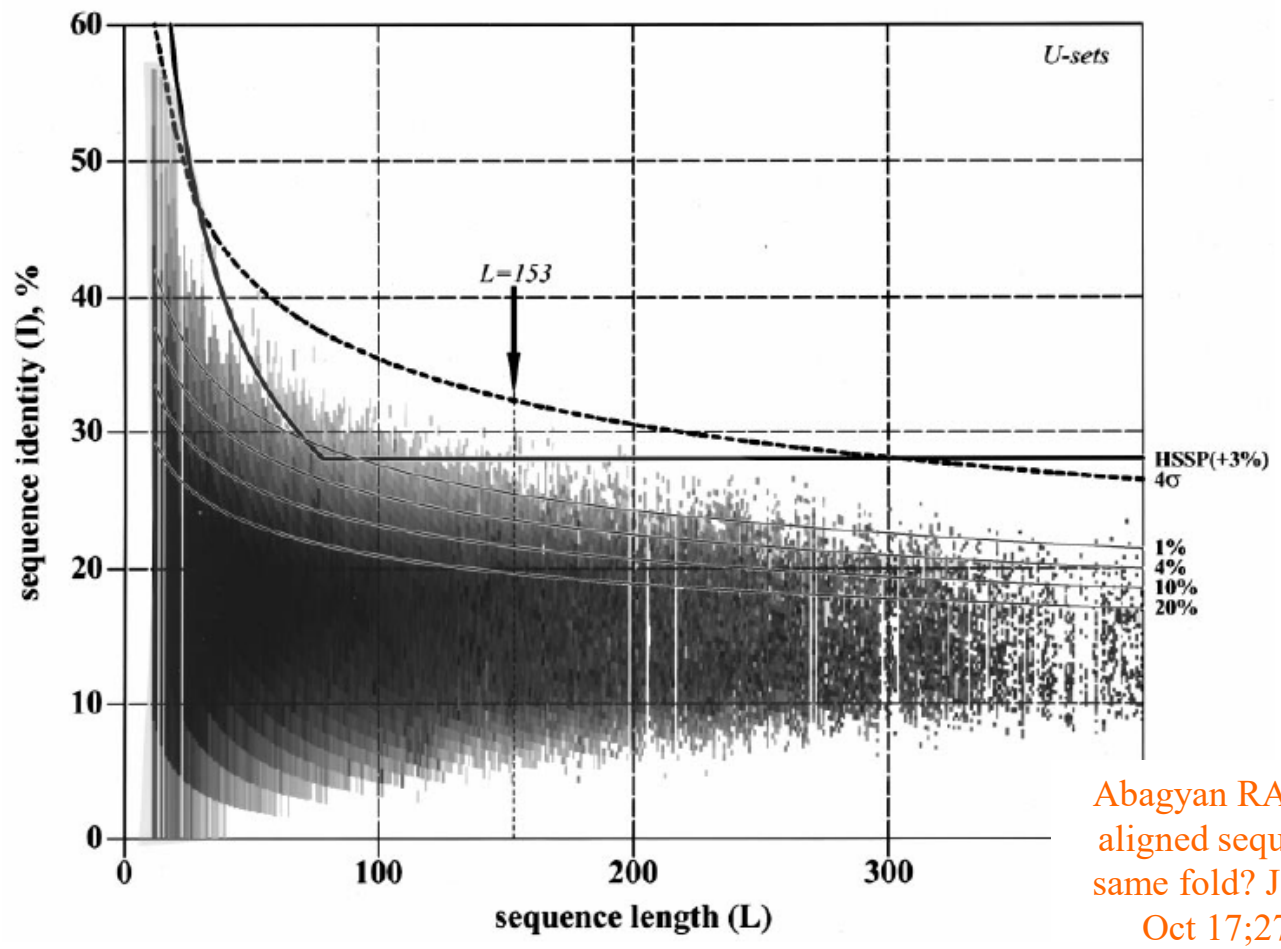
Sequences producing significant alignments:	Score (bits)	E Value
gi|14193729|gb|AAK56109.1|AF332081_1 protein tyrosin phosph...	62 L	e-177
gi|126467|sp|P18433|PTRA_HUMAN Protein-tyrosine phosphatase...	62 L	e-177
gi|4506303|ref|NP_002827.1| protein tyrosine phosphatase, r...	62 L	e-176
gi|227294|prf||1701300A protein Tyr phosphatase	620	e-176
gi|18450369|ref|NP_543030.1| protein tyrosine phosphatase, ...	62 L	e-176
gi|32067|emb|CAA37447.1| tyrosine phosphatase precursor [Ho...	61 L	e-176
gi|285113|pir||JC1285 protein-tyrosine-phosphatase (EC 3.1....	619	e-176
gi|6981446|ref|NP_036895.1| protein tyrosine phosphatase, r...	61 L	e-176
gi|2098414|pdb|1YFO|A Chain A, Receptor Protein Tyrosine Ph...	61 S	e-174
gi|32313|emb|CAA38662.1| protein-tyrosine phosphatase [Homo...	61 L	e-174
gi|450583|gb|AAB04150.1| protein tyrosine phosphatase >gi|4...	605	e-172
gi|6679557|ref|NP_033006.1| protein tyrosine phosphatase, r...	60 L	e-172
gi|483922|gb|AAA17990.1| protein tyrosine phosphatase alpha	599	e-170

Thus our example sequence could be a protein tyrosine phosphatase α (PTPα)

Example alignment with PTPα

```
 Score =  632 bits (1629), Expect = e-180
 Identities = 294/302 (97%), Positives = 294/302 (97%)

Query: 1   SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASXXXXXXXR 60
           SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAAS       R
Sbjct: 202 SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR 261

Query: 61  YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE 120
           YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
Sbjct: 262 YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE 321

Query: 121 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 180
           QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
Sbjct: 322 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 381

Query: 181 VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG 240
           VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
Sbjct: 382 VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG 441

Query: 241 TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE 300
           TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
Sbjct: 442 TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE 501
```

Twilight zone



Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol. 1997 Oct 17;273(1):355-68

Protein function prediction

```
SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
```

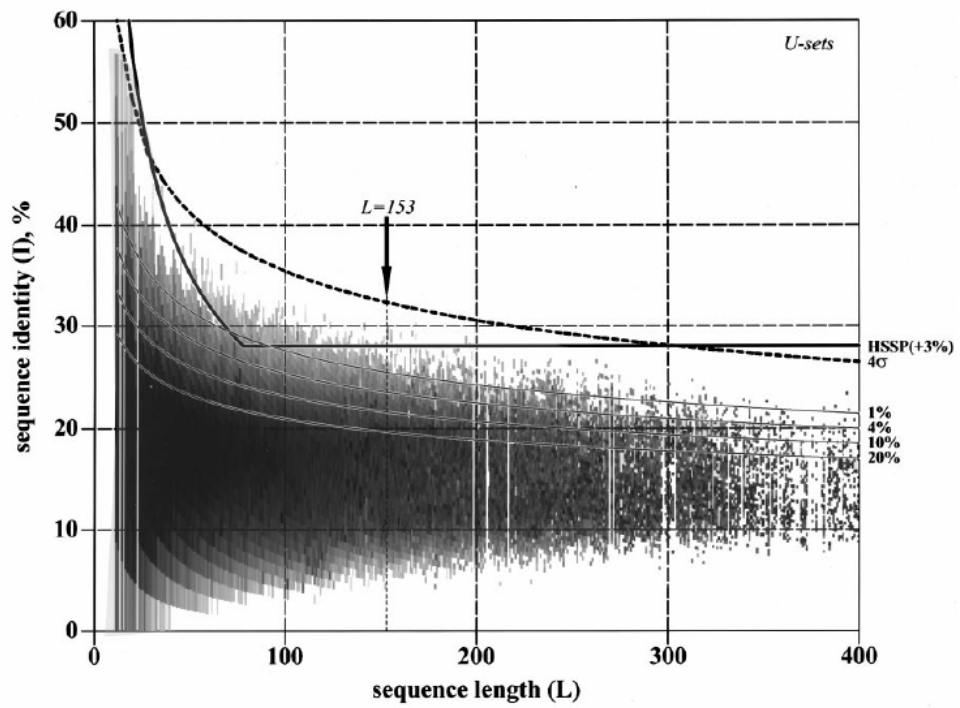
Doppelganger effect

Seq similarity to known proteins high ⇒ easy

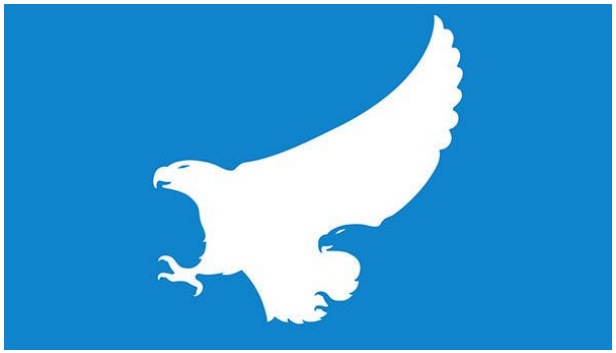
Seq similarity is low (~30%) ⇒ error prone

Seq similarity is very low ⇒ really hard

Is it possible to make more reliable protein function prediction in the twilight zone?



Inspiration



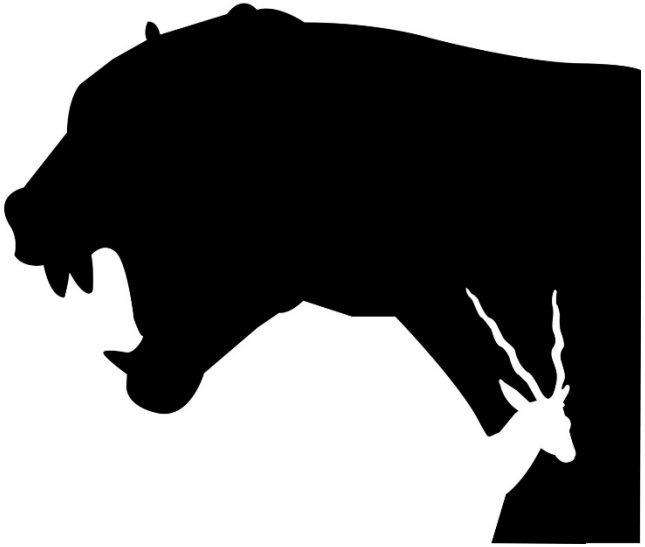
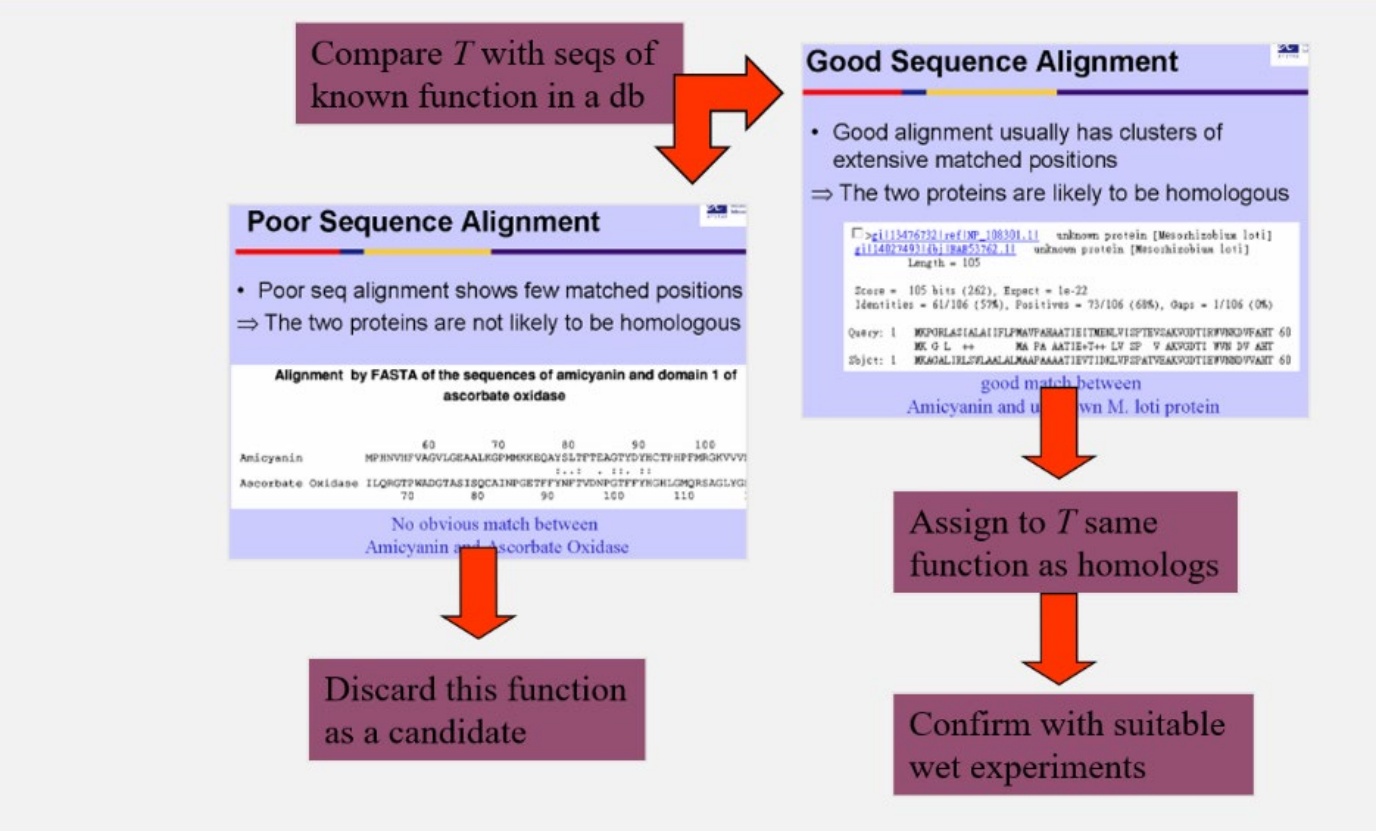


Image credit: ArtStation, Peter Hermes Furian, & TEMPLATE.NET

Guilt by association



There is no good alignment in the twilight zone.
Can poor alignments be used instead? How?

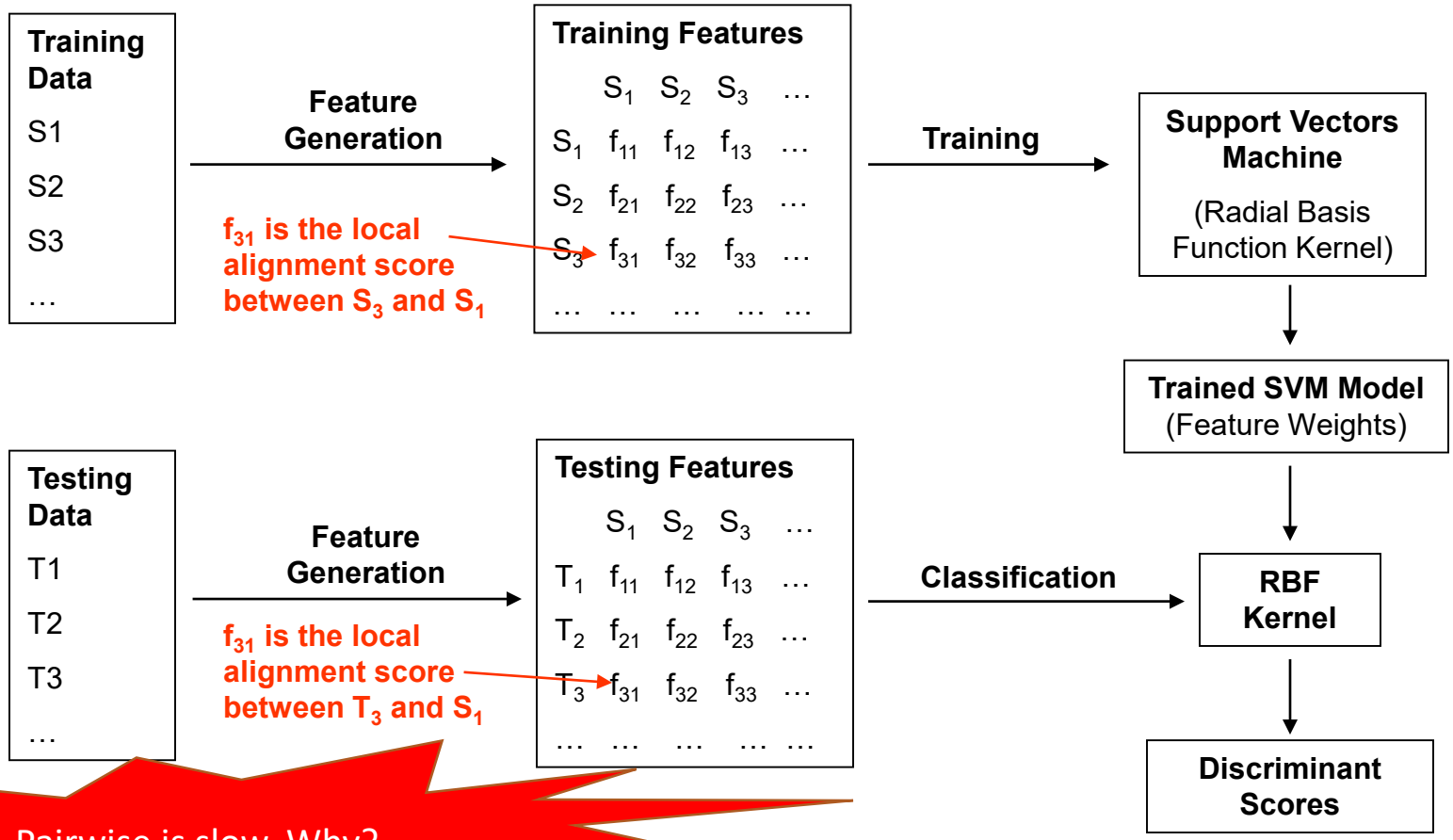
Similarities of dissimilarities

The diff betw any apple to orange / banana / mango / etc. are mostly same as the diff betw any other apple with that orange / banana / mango / etc.

The diff betw a mysterious fruit X to orange / banana / mango / etc. are mostly same as the diff betw an apple to orange / banana /mango / etc.

$\Rightarrow$The fruit X is likely an apple

SVM-Pairwise framework

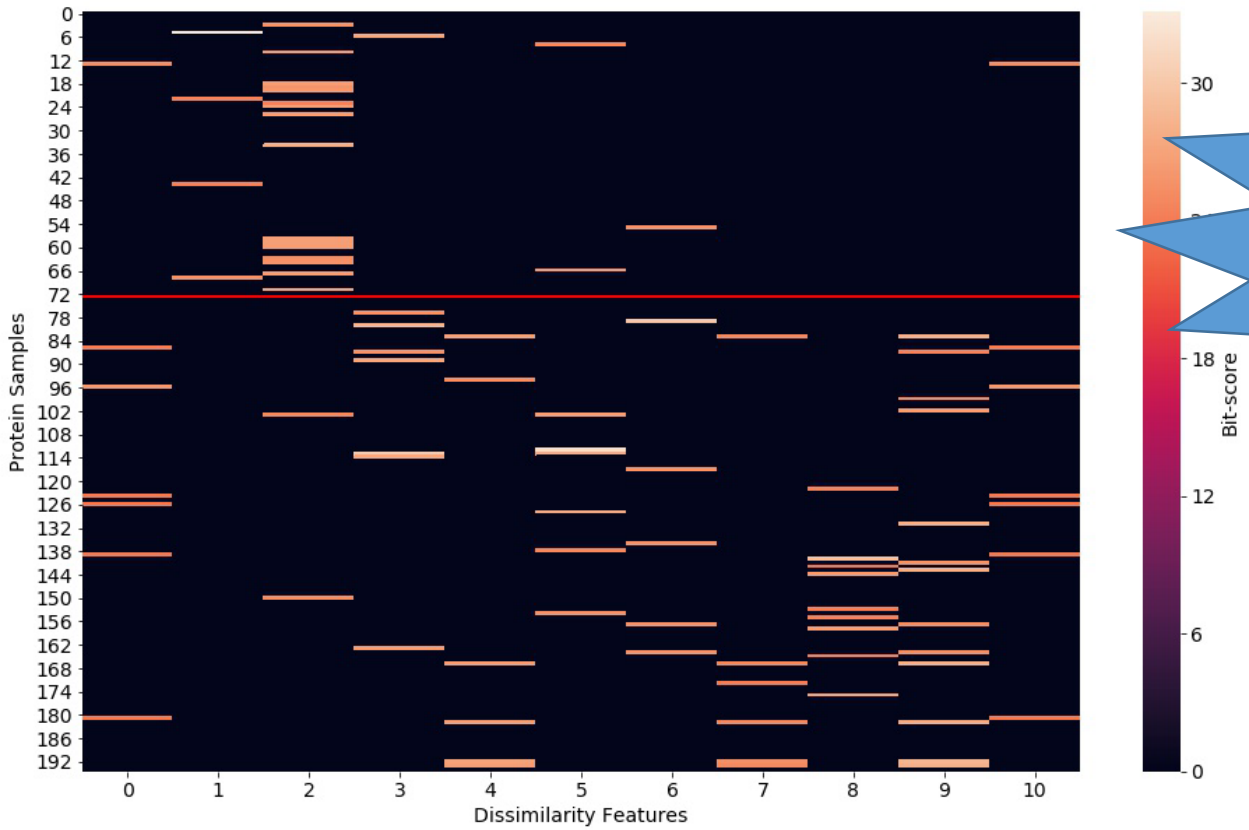


SVM-Pairwise is slow. Why?

Image credit: Kenny Chua

HeatMap of dissimilarity features

Extracted from EnsembleFam by Neamul Kabir



Note: Bitscore of proteins in the same family is ~200

There is consistency in the way two proteins of the same family differ from the other families

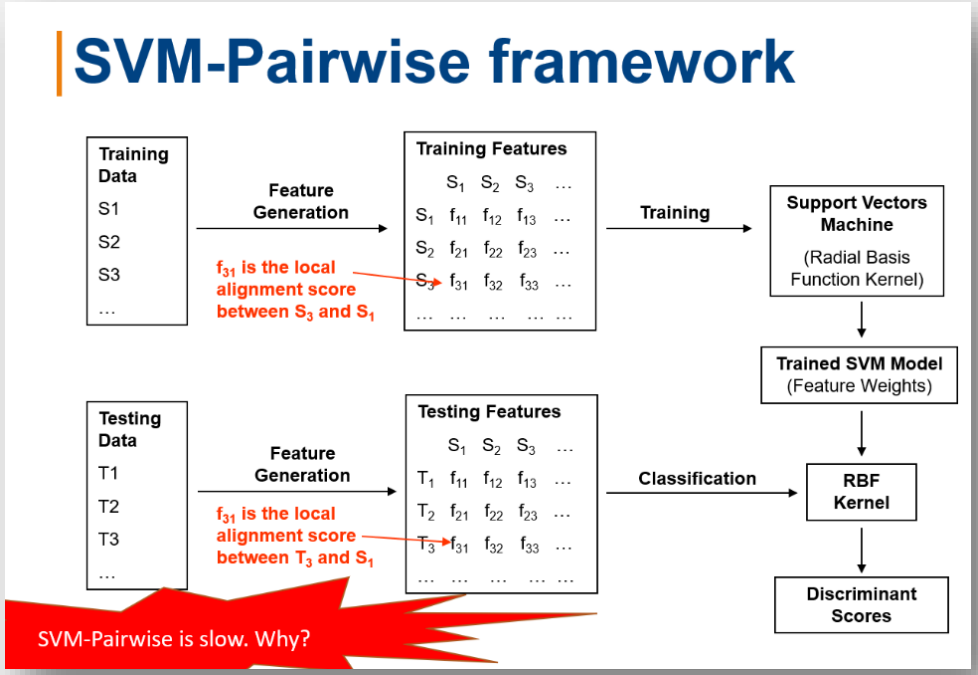
EnsembleFam

Kabir & Wong, *BMC Bioinformatics* 23:90, 2022

0 < identity <= 30							
Dataset	Method	Pred Count=1	Pred Count=2	Pred Count=3	Pred Count=4	Pred Count=5	Pred Count > 5
COG-500-1074	Ensemble Fam	**72.07**	**81.00**	**82.82**	**84.96**	**85.33**	**85.27**
	pHMM	69.54	73.75	55.51	70.62	70.85	73.55
	DeepFam	57.14	54.52	49.90	46.92	43.64	35.94
COG-250-1796	Ensemble Fam	72.84	**77.07**	**81.02**	**82.14**	**84.66**	**86.45**
	pHMM	**75.39**	73.82	73.84	71.02	67.44	72.43
	DeepFam	32.44	32.54	30.24	29.53	30.02	28.68
COG-100-2892	Ensemble Fam	**75.24**	**79.55**	**81.21**	**80.63**	**82.05**	**88.95**
	pHMM	63.44	59.69	53.45	48.16	47.42	57.57
	DeepFam	27.30	26.13	25.54	27.62	24.83	25.36

Exercise #3

SVM-Pairwise runs slowly



EnsembleFam uses the same framework but runs much faster. How does it achieve this feat?

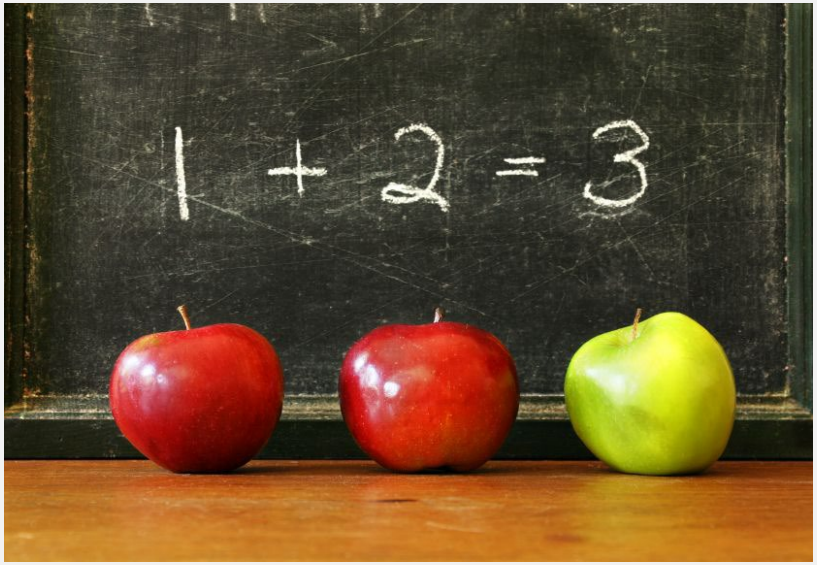
Image credit: Kenny Chua

Answer

The diff betw any apple to orange / banana / mango / etc. are mostly same as the diff betw any other apple with that orange / banana / mango / etc.

So, no need to keep the diff for 10k apples ☺

Intentionally left blank


1 + 2 = 3

What have we learned?

Doppelgangers inflate prediction performance

Stratify test set into easy vs hard instances

Similarity info is rare

Dissimilarities are aplenty but are often discarded

Guilt by association of similarity of dissimilarities is a power paradigm for classification

Expect the unexpected

There are always novel classes not modelled

References

Background:

Kabir & Wong. EnsembleFam: Towards more accurate protein family prediction in the twilight zone. *BMC Bioinformatics* 23:90, 2022

Abagyan & Batalov. Do aligned sequences share the same fold? *J Mol Biol.* 273(1):355-68, 1997

Homework:

Seo et al. DeepFam: Deep learning-based alignment-free method for protein family modelling and prediction. **Bioinformatics** 34(13):i254-i262, 2018