Insight + logic
= elegant solutions

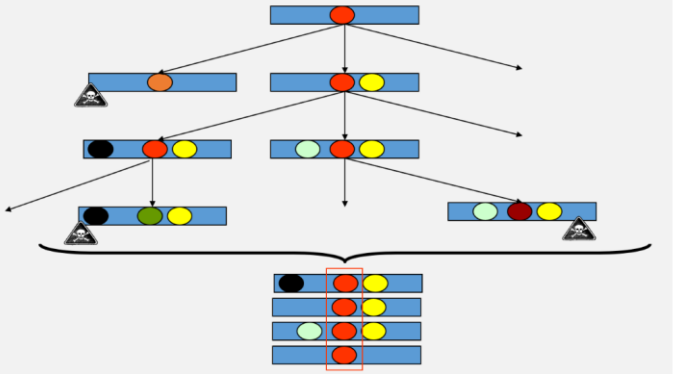
WONG Limsoon

Illuminating the twilight zone of protein function prediction

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

A standard postulate based on evolution

In the course of evolution…



Evolution takes time …

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?

Two proteins (not) inheriting their function from a common ancestor (do not) have similar amino acid sequences

Guilt by association

Compare *T* with seqs of known function in a db

Poor Sequence Alignment

- Poor seq alignment shows few matched positions

⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```
                    60        70        80        90       100
Amicyanin           MPHNVHFVAGVLGEAALKGPMMKKEQAYSLPFTEAGTYDYHCTPHPFMRGKVVV
                                                :.:  . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYG
                    70        80        90       100       110
```

No obvious match between
Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

```
>gi|13476732|ref|NP_108301.1|  unknown protein [Mesorhizobium loti]
gi|14027493|dbj|BAB53762.1|  unknown protein [Mesorhizobium loti]
          Length = 105

 Score =  105 bits (262), Expect = 1e-22
 Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1   MKPGRLASIALAIIFLPWAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
           MK G L  ++       MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

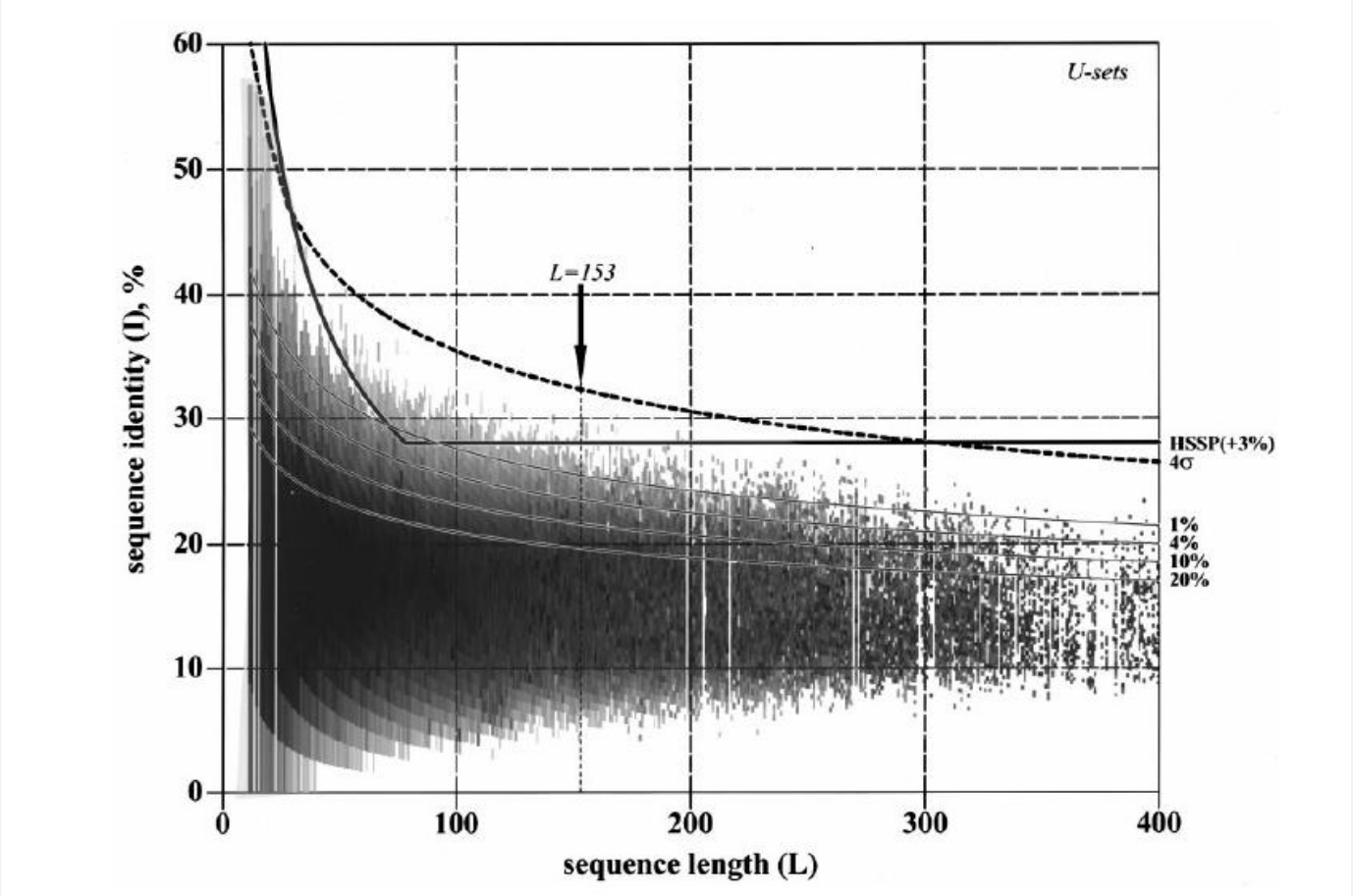
good match between
Amicyanin and unknown M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

Twilight zone: Limit of sequence similarity-based protein function assignment

So, need clever methods for the twilight zone



Abagyan RA, Batalov S. *J Mol Biol.,* 273(1):355-68, 1997

An example in the near-twilight zone

x: Human HSP70

GPLGSMSKGPAVGIDLGTTYSCVGVFQHGKVEIIANDQGNRTTPSYVAFTDT…

y: Human actin-1 (37% similarity)

MDDDIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDS…

Other families

a: B. taurus serum albumin (35%), b: E. coli leucine binding protein (35%), c: human 26S proteasome unit 7 (37%), d: A. olearius peptide chain release factor 1 (36%)

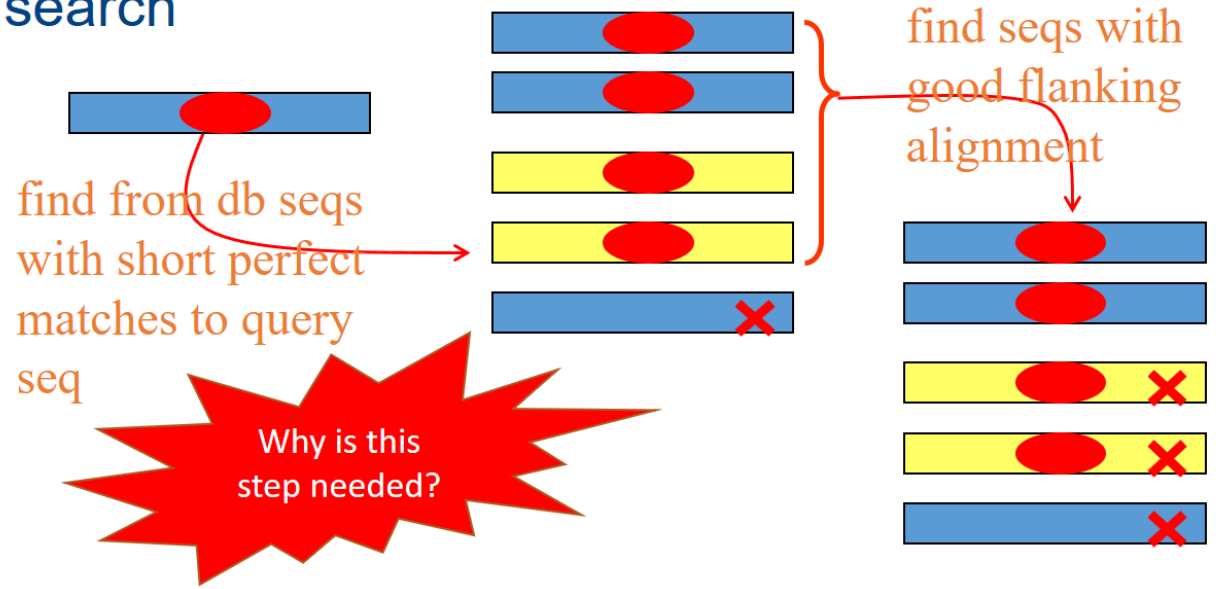
In fact, homology search tools are optimized to skip low-similarity proteins

A twilight-zone protein is low similarity ⇒ get nothing back!

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



Insight: Similarities of dissimilarities

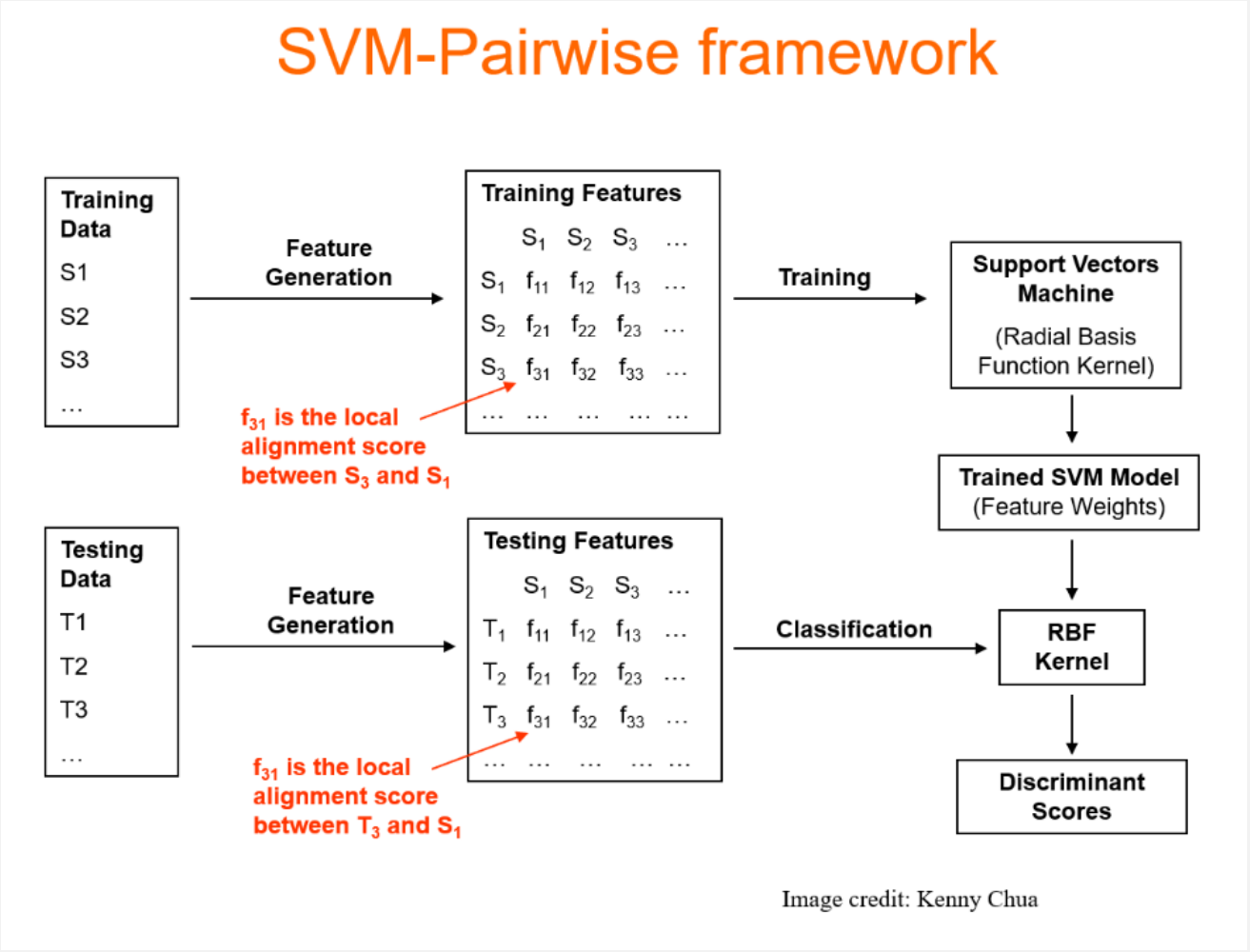
The differences between any apple to orange / banana / mango / etc. are mostly same as the differences between any other apple with that orange / banana / mango / etc.

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

$\Rightarrow$The fruit X is likely an apple

EnsembleFam uses low-/dis-similarity information discarded by other methods!

Inspired by SVM-pairwise



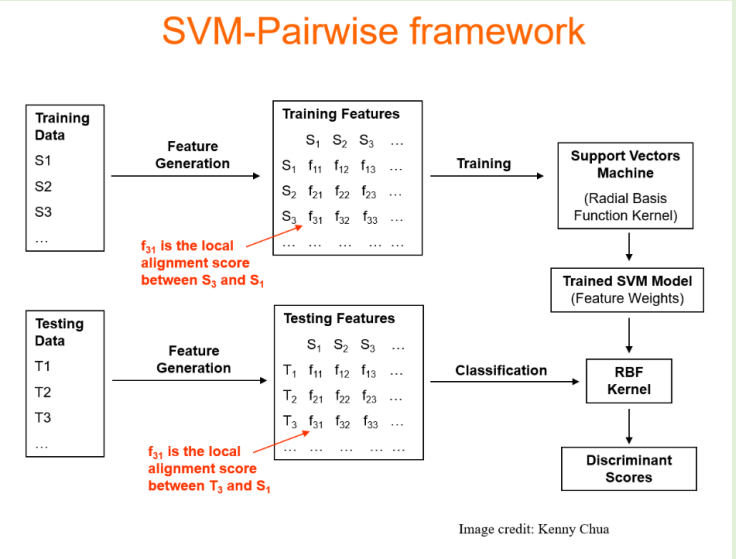
Exercise

SVM-Pairwise uses each training seq to define a feature

*COG-500-1074 has > 500 * 1074 = 537000 seqs*

⇒ *Big SVM-Pairwise model, >537000 features*

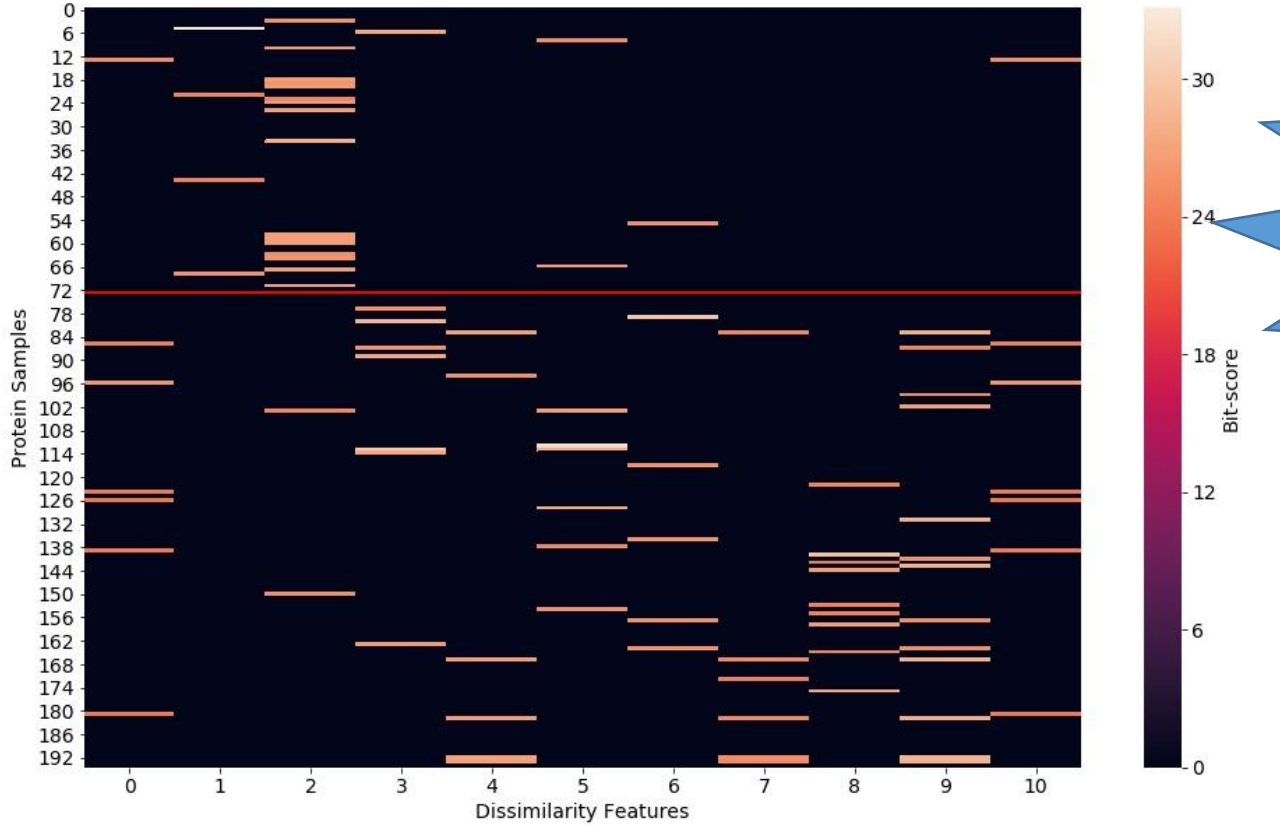
⇒ *Inefficient to compute feature vector of a test instance & make prediction for it*



Are >537000 features really needed?

How can SVM-Pairwise be made to run 100x faster?

Heatmap of dissimilarity features
Extracted from EnsembleFam



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

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

Design of EnsembleFam

One ensemble per protein family

Each ensemble has 3 base SVMs

Base SVMs use diff combinations of similarity & dissimilarity features

Ensemble decides (in vs not-in target family) by a majority vote

Feature group #1 (Dissimilarities)
Best BLAST scores from each non-target class (only 10 ref proteins used per class)

Feature group #2 (Dissimilarities)
pHMM scores from each pFam family

Feature #3 (Similarity)
Best BLAST score from target class

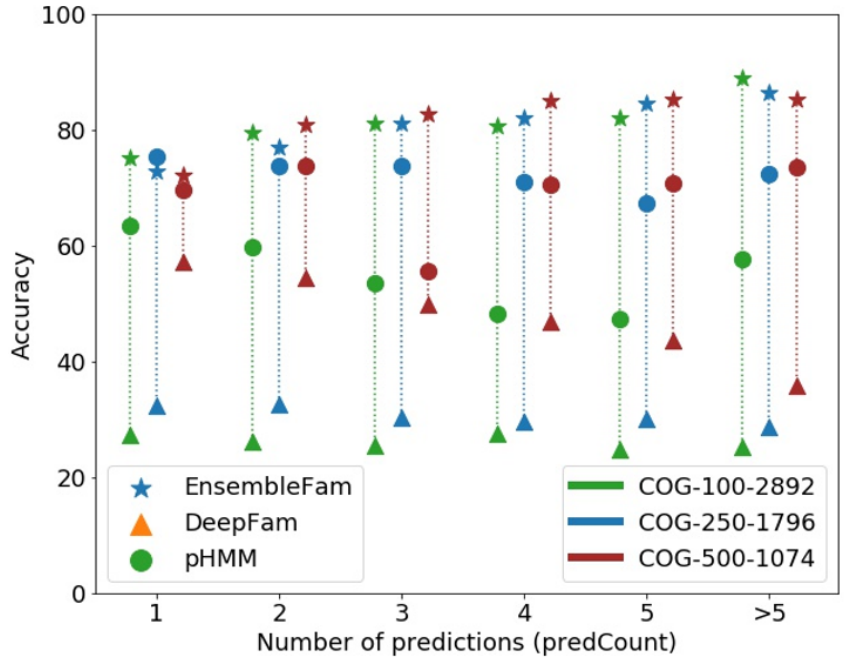
Kabir & Wong, *BMC Bioinformatics*, 23:90, 2022

Exercise

How does EnsembleFam deal with “surprising questions” ?

How does EnsembleFam deal with proteins that have more than one function?

EnsembleFam performance in the twilight zone



0 < identity <=30

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

Contribution of dissimilarities

Method	predCount = 1	predCount = 2	predCount = 3	predCount = 4	predCount = 5	predCount > 5
Identity: $0 \leq x \leq 30$						
SVM Model 1	59.82	66.96	67.60	67.96	67.32	74.67
SVM Model 2	57.11	65.09	65.01	65.86	64.55	71.80
SVM Model 3	57.34	65.34	64.29	65.02	63.36	70.13
EnsembleFam	**72.07**	**81.00**	**82.82**	**84.96**	**85.33**	**85.27**

Base SVMs have similar performance, Not much better than e.g. DeepFam

Where does performance increment of the ensemble come from?

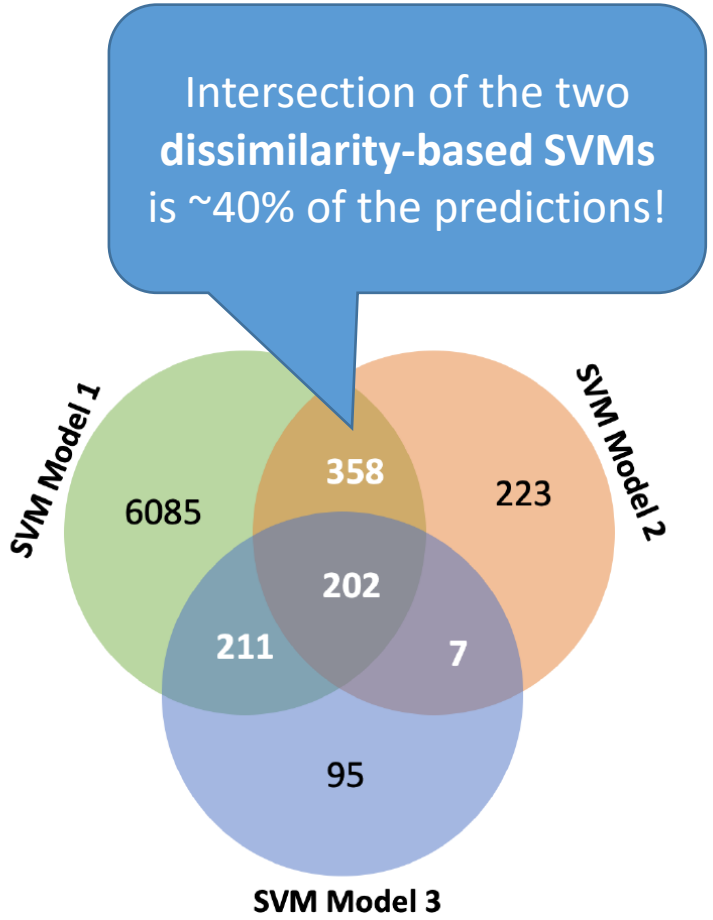


Figure 3.7: Prediction overlap of the three base classifier on the twilight zone proteins in $0 \leq$ identity ≤ 30 region. The Venn diagram is drawn based on the prediction made on twilight zone proteins of the testset of COG-500-1074 dataset. The number of predictions made by each base classifier is indicated in the figure. Numbers highlighted in white indicate overlap between at least two methods, hence predicted by EnsembleFam.

Take-home message

A lot of useful information gets overlooked

Similarity of dissimilarities

Addressing the replicability crisis

Lim et al., “A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small”, *Journal of Bioinformatics and Computational Biology*, 13(4):1550018, 2015

Poor replicability of gene selection

Low % of overlapping genes from diff expt

Prostate cancer

- Lapointe et al, 2004 vs Singh et al, 2002

Lung cancer

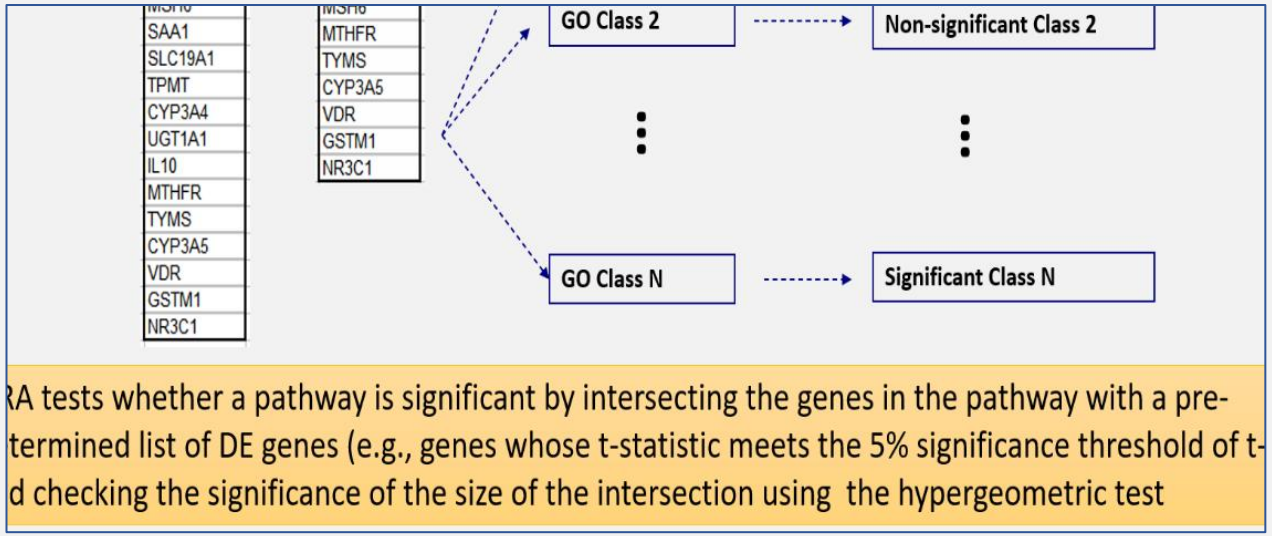
- Garber et al, 2001 vs Bhattacharjee et al, 2001

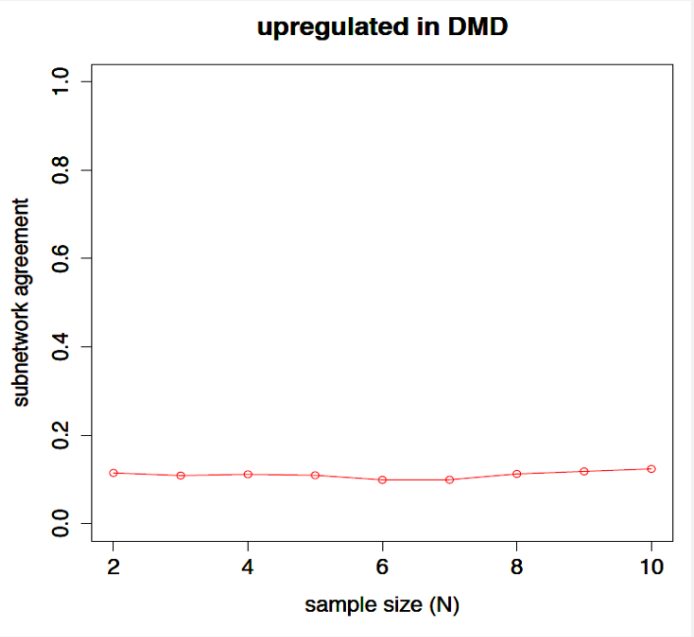
DMD

- Haslett et al, 2002 vs Pescatori et al, 2007

Datasets	DEG	POG
Prostate Cancer		
	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer		
	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD		
	Top 10	0.20
	Top 50	0.42
	Top100	0.54

We thought pathways would help but …





DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

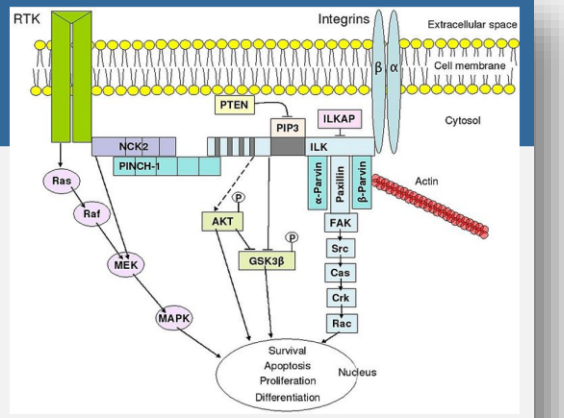
Pathway data
- PathwayAPI, Soh et al., 2010

Insight: Why ORA does not work well

Issue #1 with ORA

Its null hypothesis says "Genes in the given pathway behaves no differently from randomly chosen gene sets of the same size"

This null hypothesis is false
⇒ Lots of false positives



A biological pathway is a chain of actions of molecules in cell leading to a change in cell

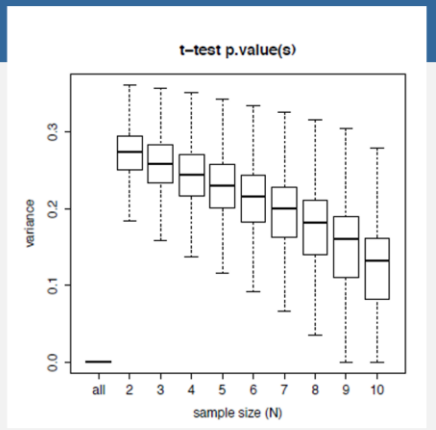
⇒ Behavour of genes in a pathway is more coordinated than random ones

Issue #2 with ORA

It relies on a pre-determined list of DEGs

This list is sensitive to the test statistic used and to the significance threshold used

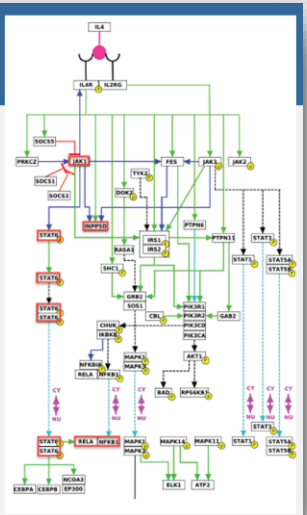
This list is unstable regardless of the threshold used when sample size is small



Issue #3 with ORA

It tests whether the entire pathway is significantly differentially expressed

If only a branch of the pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathways can dilute the signal from that branch



ORA-Paired: Paired test and new null hypothesis

Let g_i be a gene in a pathway P

Let p_j be a patient

Let q_k be a normal

Let $\Delta_{i,j,k} = Expr(g_i,p_j) – Expr(g_i,q_k)$

Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

Issue #1 is solved

Null hypothesis is "Pathway P is irrelevant to the difference between patients and normals, and the genes in P behave similarly in patients and normals"

Issue #2 is solved

No need pre-determined list of DE genes

Issue #3 is unsolved

Assume absence of batch effects

Exercise

Let g_i be a gene in a pathway P

Let p_j be a patient

Let q_k be a normal

Let $\Delta_{i,j,k} = Expr(g_i,p_j) – Expr(g_i,q_k)$

Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

How many $\Delta_{i,j,k}$ are there?

*|patients| * |normals| * |genes in P|*

Does this mean sample size now larger?

Does this mean more degrees of freedom?

Testing the null hypothesis "Pathway P is irrelevant to the difference between patients and normals and so, the genes in P behave similarly in patients and normals"

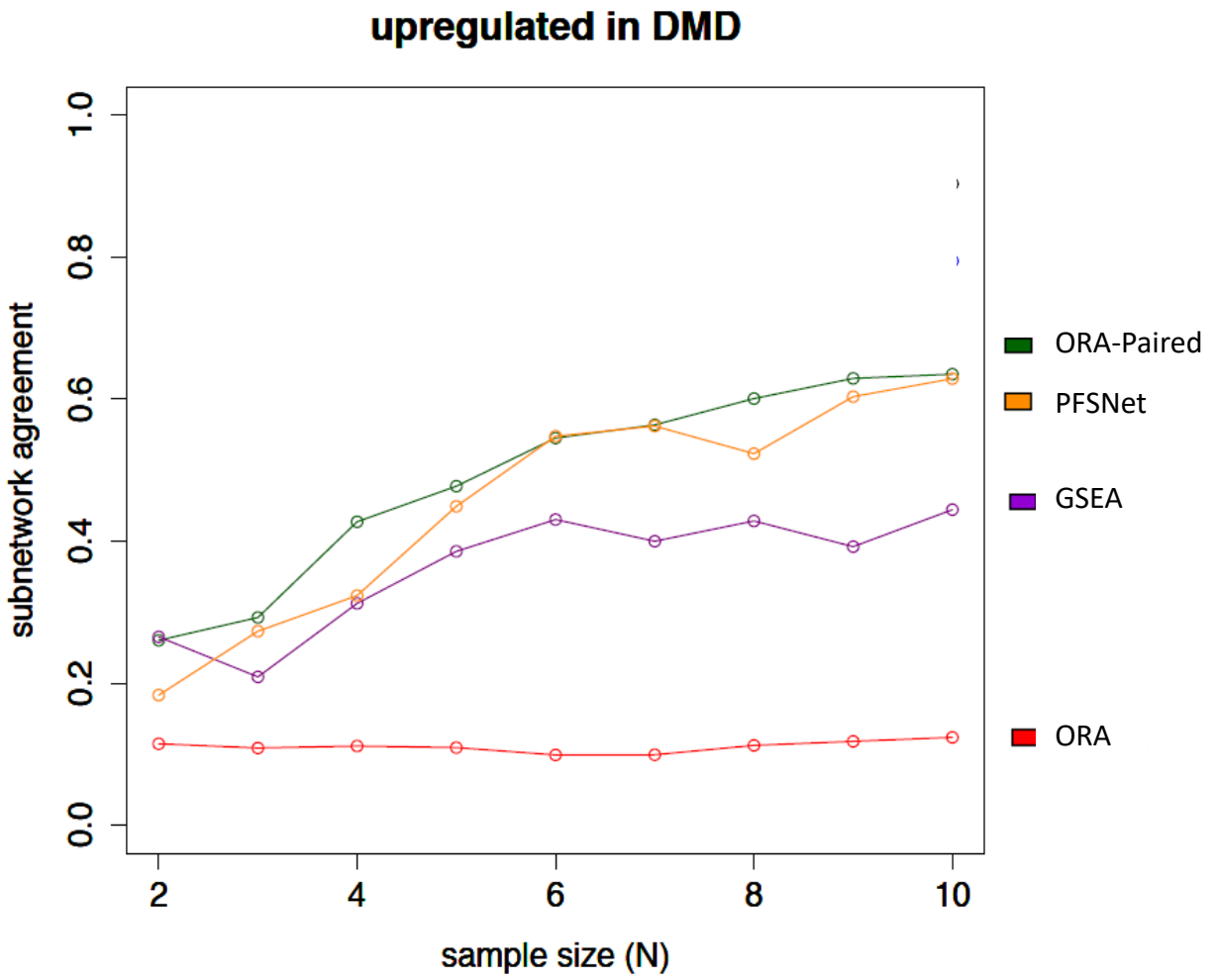
Method #1

t-test with the right degrees of freedom …

Method #2

By the null hypothesis, a dataset & its class-label permutations are exchangeable …

Better, but not super-duper good



NEA-Paired: Paired test on subnetworks

Given a pathway P

Let each node and its immediate neighbourhood in P be a subnetwork

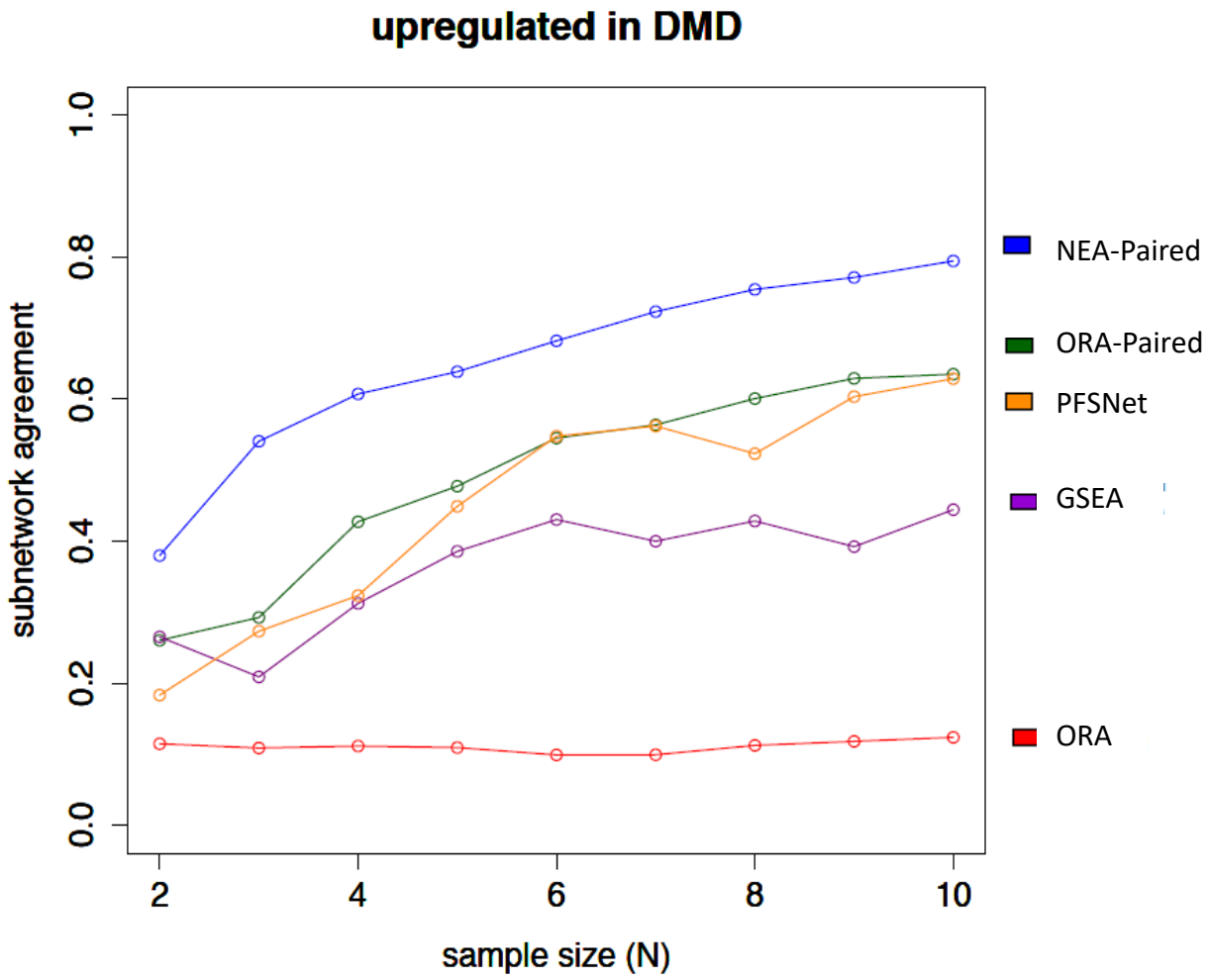
Apply ORA-Paired on each subnetwork individually

Issues #1 & #2 are solved as per ORA-Paired

Issue #3 is partly solved

Testing subnetworks instead of whole pathways

Much better performance



Take-home message

Make effort to understand the domain

A little domain insight goes a really long way

Presentation & discussion on …

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 8, AUGUST 2000

A Bayesian Computer Vision System for Modeling Human Interactions

Nuria M. Oliver, Barbara Rosario, and Alex P. Pentland, *Senior Member*, *IEEE*

WLS's comments on …

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 8, AUGUST 2000

A Bayesian Computer Vision System for Modeling Human Interactions

Nuria M. Oliver, Barbara Rosario, and Alex P. Pentland, *Senior Member*, *IEEE*