For written notes on this lecture, please read chapter 19 of *The Practical Bioinformatician* and *Hawkins & Kihara, JBCB 5(1):1-30, 2007*

CS2220: Introduction to Computational Biology

Lecture 6: Sequence Homology Interpretation

Limsoon Wong

Plan

- **Recap of sequence alignment**
- **Guilt by association**
- **Active site/domain discovery**
- **What if no homology of known function is found?**
  - Genome phylogenetic profiling
  - SVM-Pairwise
  - Protein-protein interactions
- **Key mutation site discovery**

Very Brief Recap of Sequence Comparison/Alignment

Motivations for Sequence Comparison

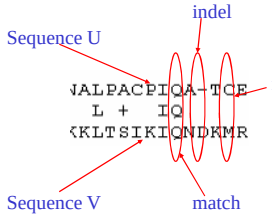
- **DNA is blue print for living organisms**

⇒ **Evolution is related to changes in DNA**

⇒ **By comparing DNA sequences we can infer evolutionary relationships between the sequences w/o knowledge of the evolutionary events themselves**

- **Foundation for inferring function, active site, and key mutations**

Sequence Alignment



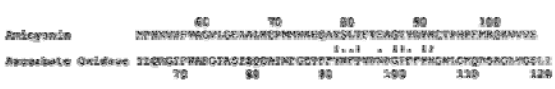
- **Key aspect of seq comparison is seq alignment**
- **A seq alignment maximizes the number of positions that are in agreement in two sequences**

Sequence Alignment: Poor Example

- **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



No obvious match between Amicyanin and Ascorbate Oxidase

Sequence Alignment: Good Example

- **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   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRVVNKDVPAHT 60
            MK G L  ++       MA PA AATIE+T++ LV SP  V AKVGDTI VVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between
Amicyanin and unknown M. loti protein

Multiple Alignment: An Example

- **Multiple seq alignment maximizes number of positions in agreement across several seqs**
- **seqs belonging to same "family" usually have more conserved positions in a multiple seq alignment**

```
gi|126467|      FHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAMLD
gi|2499753      FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|462550|      YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSMLQ
gi|2499751      FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906      FQFTAWPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAMLE
gi|126471|      LHFTSWPDFGVPFTPIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMMA
gi|548626|      FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|131570|      FHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLVVHCSAGAGRTGCFIVIDIMLD
gi|2144715      FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
                 ..* *** ***      . *              .. ***** * ****... ** ..
```

Conserved sites

Application of Sequence Comparison: Guilt-by-Association

A protein is a ...

- **A protein is a large complex molecule made up of one or more chains of amino acids**
- **Protein performs a wide variety of activities in the cell**

Function Assignment to Protein Sequence

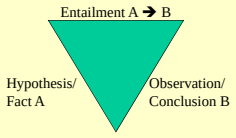
```
SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT
```

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

Invariant and Abductive Reasoning

- **Function is determined by 3D struct of protein & environment protein is in**
- **Constraints imposed by 3D struct & environment give rise to "invariant" properties observed in proteins having the ancestor with that function**

⇒ **Abductive reasoning**
- If those invariant properties are seen in a protein, then the protein is homolog of this protein



⇒ **"Guilt by association"**

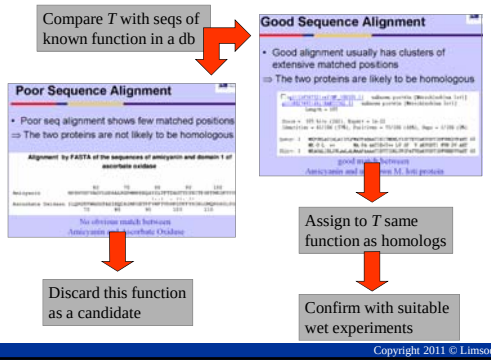
13

Guilt-by-Association

- **Compare the target sequence T with sequences S_1, …, S_n of known function in a database**
- **Determine which ones amongst S_1, …, S_n are the mostly likely homologs of T**
- **Then assign to T the same function as these homologs**
- **Finally, confirm with suitable wet experiments**

14

Guilt-by-Association

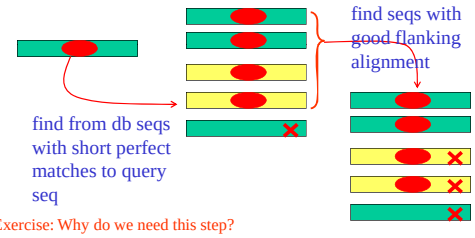


15

BLAST: How It Works

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

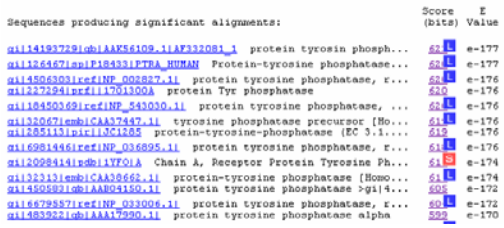
- **BLAST is one of the most popular tool for doing "guilt-by-association" sequence homology search**



Exercise: Why do we need this step?

16

Homologs obtained by BLAST



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

17

Example Alignment with PTPα

18

Guilt-by-Association: Caveats

- **Ensure that the effect of database size has been accounted for**
- **Ensure that the function of the homology is not derived via invalid "transitive assignment"**
- **Ensure that the target sequence has all the key features associated with the function, e.g., active site and/or domain**

19

Law of Large Numbers

- Suppose you are in a room with 365 other people
- Q: What is the prob that a specific person in the room has the same birthday as you?
- A: 1/365 = 0.3%
- Q: What is the prob that there is a person in the room having the same birthday as you?
- A: $1 - (364/365)^{365} = 63\%$
- Q: What is the prob that there are two persons in the room having the same birthday?
- A: 100%

20

Interpretation of P-value

- Seq. comparison progs, e.g. BLAST, often associate a P-value to each hit
- P-value is interpreted as prob that a random seq has an equally good alignment
- Suppose the P-value of an alignment is 10^{-6}
- If database has 10^7 seqs, then you expect $10^7 * 10^{-6} = 10$ seqs in it that give an equally good alignment

⇒ Need to correct for database size if your seq comparison prog does not do that!

Note: $P = 1 - e^{-E}$

Exercise: Name a commonly used method for correcting p-value for a situation like this

21

2.2 Review of BLAST statistics

In order to derive a way to combine the E-values reported by PSI-BLAST for the different rounds into a single figure of merit, it is useful to review how these E-values are actually calculated. If a query sequence with N amino acids is aligned against a random subject sequence of M amino acids using the Smith-Waterman (Smith and Waterman, 1981) local alignment algorithm[3] the local alignment score S, can be considered a random variable. In the absence of gaps, this random variable has been proven (Karlin and Altschul, 1990, 1993; Karlin and Dembo, 1992) to follow an extreme value distribution

$$\Pr\left[\sum_{ij} < x\right] = \exp\left(-KLMe^{-\lambda x}\right) \quad (1)$$

if the query and subject sequences are sufficiently long.[4] Here, λ and K are two parameters that depend in a known way on the scoring system, the query sequence and the random ensemble used to choose the random subject sequences.

Even in the presence of gaps, heuristic arguments and numerical studies (Altschul and Gish, 1996; Collins et al., 1988; Mott, 1992; Schaffer et al., 2001; Smith et al., 1985; Waterman and Vingron, 1994; Waterman and Vingron, 1994) confirm that the local alignment score is still distributed according to the extreme value distribution Equation (1), albeit with modified values of the two parameters k and λ. Thus, the P-value of an alignment score, or the probability of obtaining an alignment score of magnitude S or better in one alignment against a random subject sequence, is given by

$$P = 1 - \exp\left(-KLMe^{-\lambda S}\right) \quad (2)$$

BLAST and PSI-BLAST report E-values, or the expected number of random hits of score S or higher, instead of P-values. These two quantities are different since the P-value refers to the probability of having at least one hit at the prescribed significance level, while the E-value quantifies the expected number of these hits. The general relationship between these two quantities for random variables with exponential tails is the Poisson formula

$$P = 1 - e^{-E} \quad (3)$$

By comparing with Equation (2) we find

$$E_{\mathrm{seq}} = -\ln(1-P) \approx KLMe^{-\lambda S} \quad (4)$$

It is important to note that this is the E-value for *one single sequence* comparison, which we indicate by the subscript 'seq'. The convenient property of E-values is that the Bonferroni correction for repeating the alignment for every sequence in a database of N amino acids and thus roughly N/M sequences, just manifests itself in a multiplication of the expected number of hits for a single sequence by the number of sequences, i.e.

$$E_{\mathrm{db}} = \frac{N}{M} E_{\mathrm{seq}} = KLNe^{-\lambda S} \quad (5)$$

22

Lightning Does Strike Twice!

- Roy Sullivan, a former park ranger from Virgina, was struck by lightning 7 times
  - 1942 (lost big-toe nail)
  - 1969 (lost eyebrows)
  - 1970 (left shoulder seared)
  - 1972 (hair set on fire)
  - 1973 (hair set on fire & legs seared)
  - 1976 (ankle injured)
  - 1977 (chest & stomach burned)
- September 1983, he committed suicide

Cartoon: Ron Hipschman
Data: David Hand

23

Effect of Seq Compositional Bias

- One fourth of all residues in protein seqs occur in regions with biased amino acid composition
- Alignments of two such regions achieves high score purely due to segment composition

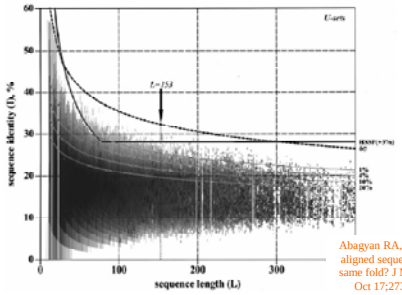
⇒ While it is worth noting that two proteins contain similar low complexity regions, they are best excluded when constructing alignments

- E.g., by default, BLAST employs the SEG algo to filter low complexity regions from proteins before executing a search

Source: NCBI

24

Effect of Sequence Length



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

25

Examples of Invalid Function Assignment:

The IMP Dehydrogenases (IMPDH)

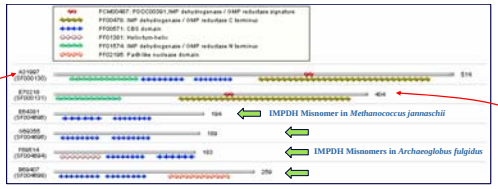
10 entries were found

ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept
NP_001181857	Methanococcus jannaschii	B64581 conserved hypothetical protein MJ0633	Y633_METJA Hypothetical protein MJ0633	4179288 inosine-5'-monophosphate dehydrogenase (guaB) NP_247627 inosine-5'-monophosphate dehydrogenase (guaB)
NP_001187788	Archaeoglobus fulgidus	G69332 MJ0633 homolog AF0847 ALT_NAME2: inosine-monophosphate dehydrogenase (guaB-1) homolog [numcmon]	O28461 INOSINE MONOPHOSPHATE DEHYDROGENASE (IMAB-1)	g2648724 inosine monophosphate dehydrogenase (guaB-1) NP_069021 inosine monophosphate dehydrogenase (guaB-1)
NP_001188267	Archaeoglobus fulgidus	F69258 yhcF homolog 2 ALT_NAME2: inosine-monophosphate dehydrogenase (guaB-2) homolog [numcmon]	O28142 INOSINE MONOPHOSPHATE DEHYDROGENASE (IMAB-2)	g2648482 inosine monophosphate dehydrogenase (guaB-2) NP_070045 inosine monophosphate dehydrogenase (guaB-2)
NP_001189697	Archae			ophosphate ase nosphosphate ase
NP_001197776	Thermo			inophosphate d protein nophosphate d protein
NP_001434709	Methanothermobacter thermautotrophicus		O27284 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	dehydrogenase related protein V NP_276234 inosine-5'-monophosphate dehydrogenase related protein V
NP_001434811	Methanothermobacter thermautotrophicus	D69502 MJ1232 protein homolog MTH126 ALT_NAME2: inosine-5'-monophosphate dehydrogenase related protein VII [numcmon]	O26529 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII	g2622166 inosine-5'-monophosphate dehydrogenase related protein VII NP_275289 inosine-5'-monophosphate dehydrogenase related protein VII
NP_001434837	Methanothermobacter thermautotrophicus	H69258 MJ1232-related protein MTH992 ALT_NAME2: inosine-5'-monophosphate dehydrogenase related protein IX [numcmon]	O27071 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN IX	g2622003 inosine-5'-monophosphate dehydrogenase related protein IX NP_276147 inosine-5'-monophosphate dehydrogenase related protein IX
NP_001434969	Methanothermobacter thermautotrophicus	B69077 yhcF homolog 2 ALT_NAME2: inosine-monophosphate dehydrogenase related protein X [numcmon]	O27016 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN X	g2622007 inosine-5'-monophosphate dehydrogenase related protein X NP_276087 inosine-5'-monophosphate dehydrogenase related protein X

A partial list of IMPdehydrogenase misnomers in complete genomes remaining in some public databases

26

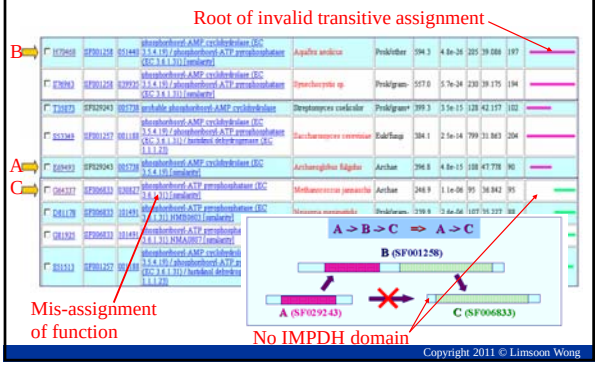
IMPDH Domain Structure



- **Typical IMPDHs have 2 IMPDH domains that form the catalytic core and 2 CBS domains.**
- **A less common but functional IMPDH (E70218) lacks the CBS domains.**
- **Misnomers show similarity to the CBS domains**

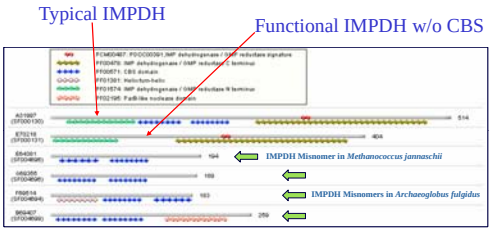
27

Invalid Transitive Assignment



28

Emerging Pattern



- **Most IMPDHs have 2 IMPDH and 2 CBS domains**
- **Some IMPDH (E70218) lacks CBS domains**

⇒ **IMPDH domain is the emerging pattern**

Application of Sequence Comparison: Active Site/Domain Discovery

30

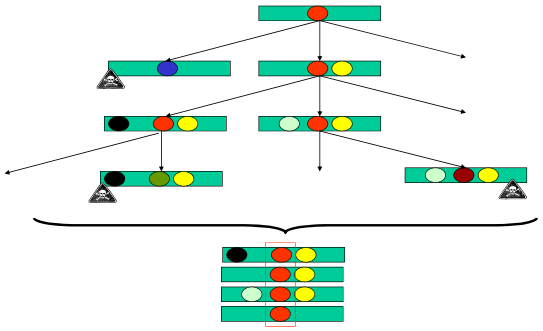
Discover Active Site and/or Domain

- **How to discover the active site and/or domain of a function in the first place?**
  - Multiple alignment of homologous seqs
  - Determine conserved positions
  - ⇒ Emerging patterns relative to background
  - ⇒ Candidate active sites and/or domains
- **Easier if sequences of distance homologs are used**

Exercise: Why?

31

In the course of evolution…



32

Multiple Alignment of PTPs

```
gi|126467|      FHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAMLD
gi|2499753      FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|462550|      YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSMLQ
gi|2499751      FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906      FQFTAWPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAMLE
gi|126471|      LHFTSWPDFGVPFTPIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMMA
gi|548626|      FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|131570|      FHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLVVHCSAGAGRTGCFIVIDIMLD
gi|2144715      FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
                 ..* *** ***      . *               ..****** **** ... ** ..
```

- **Notice the PTPs agree with each other on some positions more than other positions**
- **These positions are more impt wrt PTPs**
- **Else they wouldn't be conserved by evolution**

⇒ **They are candidate active sites**

Guilt-by-Association: What if no homolog of known function is found?

34

What if there is no useful seq homolog?

- **Guilt by other types of association!**
  - Domain modeling (e.g., HMMPFAM)
  - ✓ Similarity of phylogenetic profiles
  - ✓ Similarity of dissimilarities (e.g., SVM-PAIRWISE)
  - Similarity of subcellular co-localization & other physico-chemico properties(e.g., PROTFUN)
  - Similarity of gene expression profiles
  - ✓ Similarity of protein-protein interaction partners
  - …
  - Fusion of multiple types of info

35

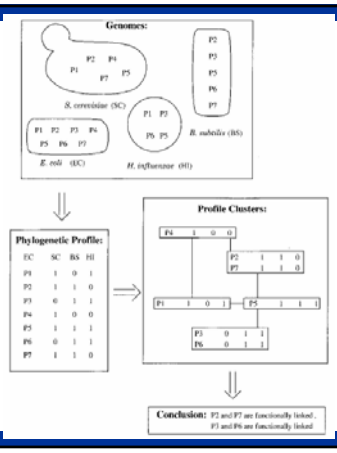
Phylogenetic Profiling

Pellegrini et al., *PNAS*, 96:4285--4288, 1999

- **Gene (and hence proteins) with identical patterns of occurrence across phyla tend to function together**

⇒ **Even if no homolog with known function is available, it is still possible to infer function of a protein**

36



Phylogenetic Profiling: How it Works

Phylogenetic Profiling: P-value

The probability of observing by chance z occurrences of genes X and Y in a set of N lineages, given that X occurs in x lineages and Y in y lineages is

$$P(z|N,x,y) = \frac{w_z * w_C}{W}$$

where

$$w_z = \binom{N}{z}$$

$$w_C = \binom{N-z}{x-z} * \binom{N-x}{y-z}$$

$$W = \binom{N}{x} * \binom{N}{y}$$

No. of ways to distribute z co-occurrences over N lineage's

No. of ways to distribute the remaining $x-z$ and $y-z$ occurrences over the remaining $N-z$ lineage's

No. of ways of distributing X and Y over N lineage's without restriction

Phylogenetic Profiles: Evidence

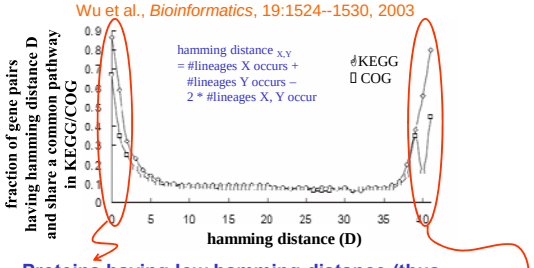
Pellegrini et al., *PNAS*, 96:4285--4288, 1999

Keyword	No. of non-homologous proteins in group	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin and Molybdenum, and molybdoterin	12	6	1
Hypothetical†	1,084	108,226	8,440

- **E. coli proteins grouped based on similar keywords in SWISS-PROT have similar phylogenetic profiles**

Phylogenetic Profiling: Evidence

Wu et al., *Bioinformatics*, 19:1524--1530, 2003



- **Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways**

Exercise: Why do proteins having high hamming distance also have this behaviour?

Guilt by Association of Dissimilarities

Differences of "unknown" to other fruits are same as "apple" to other fruits

"unknown" is an "apple"!

	Orange$_1$	Banana$_1$	...
Apple$_1$	Color = red vs orange Skin = smooth vs rough **Size = small vs small** **Shape = round vs round**	Color = red vs yellow **Skin = smooth vs smooth** **Size = small vs small** Shape = round vs oblong	...
Orange$_2$	**Color = orange vs orange** **Skin = rough vs rough** **Size = small vs small** **Shape = round vs round**	Color = orange vs yellow **Skin = rough vs smooth** **Size = small vs small** Shape = round vs oblong	...
Unknown$_1$	**Color = red vs orange** **Skin = smooth vs rough** Size = small vs small Shape = round vs round	**Color = red vs yellow** Skin = smooth vs smooth Size = small vs small **Shape = round vs oblong**	...
...	...	...	...

SVM-Pairwise Framework

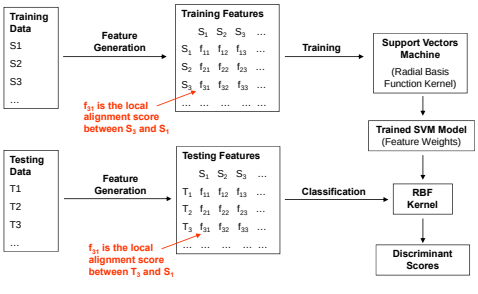
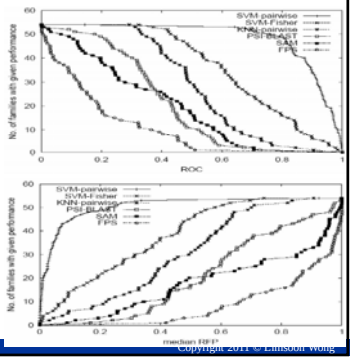


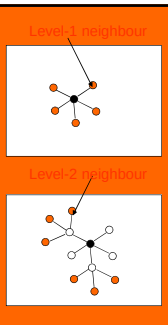
Image credit: Kenny Chua

Performance of SVM-Pairwise

- **Receiver Operating Characteristic (ROC)**
  - The area under the curve derived from plotting true positives as a function of false positives for various thresholds.
- **Rate of median False Positives (RFP)**
  - The fraction of negative test examples with a score better or equals to the median of the scores of positive test examples.

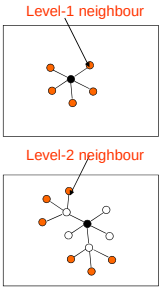


Protein Function Prediction from Protein Interactions

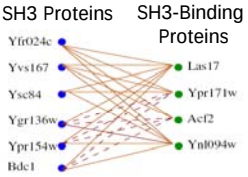


Functional Association Thru Interactions

- **Direct functional association:**
  - Interaction partners of a protein are likely to share functions w/ it
  - Proteins from the same pathways are likely to interact
- **Indirect functional association**
  - Proteins that share interaction partners with a protein may also likely to share functions w/ it
  - Proteins that have common biochemical, physical properties and/or subcellular localization are likely to bind to the same proteins

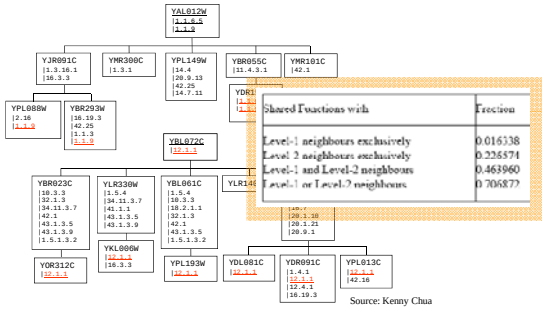


An illustrative Case of Indirect Functional Association?



- **Is *indirect functional association* plausible?**
- **Is it found often in real interaction data?**
- **Can it be used to improve protein function prediction from protein interaction data?**

Freq of Indirect Functional Association



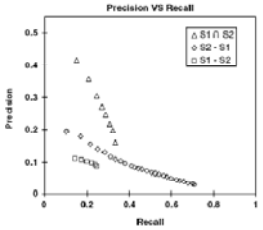
Source: Kenny Chua

Prediction Power By Majority Voting

- **Remove overlaps in level-1 and level-2 neighbours to study predictive power of "level-1 only" and "level-2 only" neighbours**
- **Sensitivity vs Precision analysis**

$$PR = \frac{\sum_i^K k_i}{\sum_i^K m_i} \qquad SN = \frac{\sum_i^K k_i}{\sum_i^K n_i}$$

- n_i is no. of fn of protein i
- m_i is no. of fn predicted for protein i
- k_i is no. of fn predicted correctly for protein i



⇒ **"level-2 only" neighbours performs better**

⇒ **L1 ∩ L2 neighbours has greatest prediction power**

Functional Similarity Estimate: Czekanowski-Dice Distance

- **Functional distance between two proteins** (Brun et al, 2003)

$$D(u,v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}$$

- N_k is the set of interacting partners of k
- X Δ Y is symmetric diff betw two sets X and Y
- Greater weight given to similarity

Is this a good measure if u and v have very diff number of neighbours?

⇒ **Similarity can be defined as**

$$S(u,v) = 1 - D(u,v) = \frac{2X}{2X + (Y + Z)}$$

Functional Similarity Estimate: FS-Weighted Measure

- **FS-weighted measure**

$$S(u,v)=\frac{2|N_u\cap N_v|}{|N_u-N_v|+2|N_u\cap N_v|}\times\frac{2|N_u\cap N_v|}{|N_v-N_u|+2|N_u\cap N_v|}$$

  - N_k is the set of interacting partners of k
  - Greater weight given to similarity

⇒ **Rewriting this as**

$$S(u,v)=\frac{2X}{2X+Y}\times\frac{2X}{2X+Z}$$

Correlation w/ Functional Similarity

- **Correlation betw functional similarity & estimates**

Neighbours	CD-Distance	FS-Weight
S_1	0.471810	0.498745
S_2	0.224705	0.298843
$S_1 \cup S_2$	0.224581	0.29629

- **Equiv measure slightly better in correlation w/ similarity for L1 & L2 neighbours**

Reliability of Expt Sources

- **Diff Expt Sources have diff reliabilities**
  - Assign reliability to an interaction based on its expt sources (Nabieva et al, 2004)
- **Reliability betw u and v computed by:**

$$r_{u,v}=1-\prod_{i\in E_{u,v}}(1-r_i)$$

  - r_i is reliability of expt source i,
  - $E_{u,v}$ is the set of expt sources in which interaction betw u and v is observed

Source	Reliability
Affinity Chromatography	0.823077
Affinity Precipitation	0.455904
Biochemical Assay	0.666667
Dosage Lethality	0.5
Purified Complex	0.891473
Reconstituted Complex	0.5
Synthetic Lethality	0.37386
Synthetic Rescue	1
Two Hybrid	0.265407

Functional Similarity Estimate: FS-Weighted Measure with Reliability

- **Take reliability into consideration when computing FS-weighted measure:**

$$S_R(u,v)=\frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_u-N_v} r_{u,w}+\sum_{w\in(N_u\cap N_v)} r_{u,w}(1-r_{v,w})\right)+2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}\times\frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_v-N_u} r_{v,w}+\sum_{w\in(N_u\cap N_v)} r_{v,w}(1-r_{u,w})\right)+2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}$$

  - N_k is the set of interacting partners of k
  - $r_{u,w}$ is reliability weight of interaction betw u and v

⇒ **Rewriting**

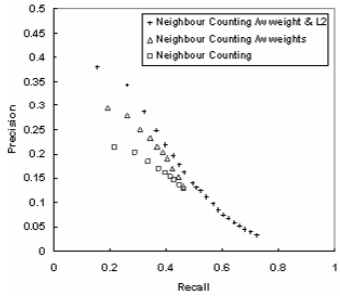
$$S(u,v)=\frac{2X}{2X+Y}\times\frac{2X}{2X+Z}$$

Integrating Reliability

- **Equiv measure shows improved correlation w/ functional similarity when reliability of interactions is considered:**

Neighbours	CD-Distance	FS-Weight	FS-Weight R
S_1	0.471810	0.498745	0.532596
S_2	0.224705	0.298843	0.375317
$S_1 \cup S_2$	0.224581	0.29629	0.363025

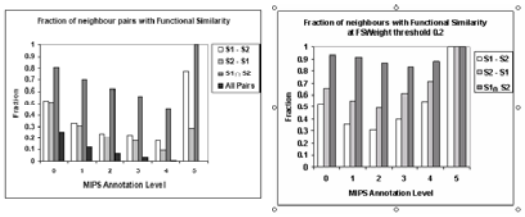
Improvement to Prediction Power by Majority Voting



Considering only neighbours w/ FS weight > 0.2

55

Improvement to Over-Rep of Functions in Neighbours



56

Use L1 & L2 Neighbours for Prediction

- **FS-weighted Average**

$$f_x(u) = \frac{1}{Z}\left[\lambda r_{\text{int}}\pi_x + \sum_{v\in N_u}\left(S_{TR}(u,v)\delta(v,x) + \sum_{w\in N_v} S_{TR}(u,w)\delta(w,x)\right)\right]$$

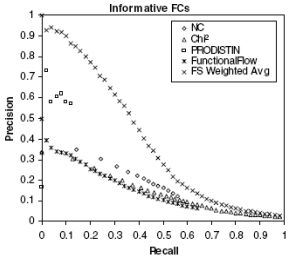
- r_{int} is fraction of all interaction pairs sharing function
- λ is weight of contribution of background freq
- $\delta(k, x) = 1$ if k has function x, 0 otherwise
- N_k is the set of interacting partners of k
- π_x is freq of function x in the dataset
- Z is sum of all weights

$$Z = 1 + \sum_{v\in N_u}\left(S_{TR}(u,v) + \sum_{w\in N_v} S_{TR}(u,w)\right)$$

57

Performance of FS-Weighted Averaging

- **LOOCV comparison with Neighbour Counting, Chi-Square, PRODISTIN**



58

About the Inventor: Chua Hon Nian

- **Chua Hon Nian**
  - PhD, NUS, 2008
  - Postdoc at Harvard & Univ of Toronto
  - 49th hottest paper in Computer Science published in 2006
  - Winner, DREAM2 challenge PPI subnetwork, 2007

Application of Sequence Comparison: Key Mutation Site Discovery

60

Identifying Key Mutation Sites

K.L.Lim et al., *JBC*, 273:28986--28993, 1998

Sequence from a typical PTP domain D2

```
>gi|000000|PTPA-D2
EEEFKKLTSIKIQNDKNRTGNLPANHKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASF
IDGYRQKDSYIASQGPLLHTIEDFWRMIWESGSSIVVMLTELEERGQEKCAQYWPSDGLV
SYGDITVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFEGWPEVGIPSDGKGMISII
AAVQKQQQQSGNHPITVHCSAGAGRTGTFCALSTVLERVKAEGILDVFQTVKSLRLQRPH
MVQTLEQYEFCYKVVQEYIDAFSDYANFK
```

- **Some PTPs have 2 PTP domains**
- **PTP domain D1 is has much more activity than PTP domain D2**
- **Why? And how do you figure that out?**

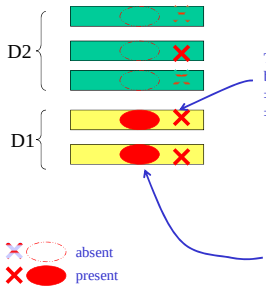
61

Emerging Patterns of PTP D1 vs D2

- **Collect example PTP D1 sequences**
- **Collect example PTP D2 sequences**
- **Make multiple alignment A1 of PTP D1**
- **Make multiple alignment A2 of PTP D2**
- **Are there positions conserved in A1 that are violated in A2?**
- **These are candidate mutations that cause PTP activity to weaken**
- **Confirm by wet experiments**

62

Emerging Patterns of PTP D1 vs D2



This site is consistently conserved in D1, but is not consistently missing in D2
⇒ it is not an EP
⇒ not a likely cause of D2's loss of function

Exercise: Why?

This site is consistently conserved in D1, but is consistently missing in D2
⇒ it is an EP
⇒ possible cause of D2's loss of function

63

Key Mutation Site: PTP D1 vs D2

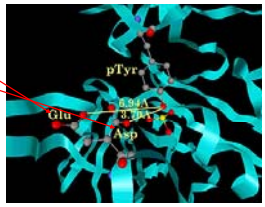
```
                    ?   !  ?          ?                    ?              ? ??
gi|00000|P  D2  QFHFHGWPEVGIPSDGKGMISIIAAVQKQQQQ-SGNHPITVHCSAGAGRTGTFCALSTVL
gi|126467|      QFHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAML
gi|2499753      QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML
gi|462550|      QYHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSML
gi|2499751      QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLI
gi|1709906  D1  QFQFTAWPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAML
gi|126471|      QLHFTSWPDFGVPFTPIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMM
gi|548626|      QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCFIVIDIML
gi|131570|      QFHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLVVHCSAGAGRTGCFIVIDIML
gi|2144715      QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLI
                * ..  **. *.*      .                . ****** ****..  .  ..
```

- **Positions marked by "!" and "?" are likely places responsible for reduced PTP activity**
  - All PTP D1 agree on them
  - All PTP D2 disagree on them

64

Key Mutation Site: PTP D1 vs D2

```
                    ?   !  ?
gi|00000|P  D2  QFHFHGWPEVGIPSDGK(
gi|126467|      QFHFTSWPDFGVPFTPI(
gi|2499753      QFHFTGWPDHGVPYHAT(
gi|462550|      QYHYTQWPDMGVPEYALI
gi|2499751      QFHFTSWPDHGVPDTTDI
gi|1709906  D1  QFQFTAWPDHGVPEHPTI
gi|126471|      QLHFTSWPDFGVPFTPI(
gi|548626|      QFHFTGWPDHGVPYHAT(
gi|131570|      QFHFTGWPDHGVPYHAT(
gi|2144715      QFHFTSWPDHGVPDTTDI
                * ..  **. *.*
```



- **Positions marked by "!" are even more likely as 3D modeling predicts they induce large distortion to structure**

65

Confirmation by Mutagenesis Expt

- **What wet experiments are needed to confirm the prediction?**
  - Mutate E → D in D2 and see if there is gain in PTP activity
  - Mutate D → E in D1 and see if there is loss in PTP activity

Exercise: Why do you need this 2-way expt?

66

About the Inventor: Prasanna Kolatkar

- **Prasanna Kolatkar**
  - Research Fellow, BIC, NUS, 1997-1999
  - Currently Group Leader at GIS

Concluding Remarks

What have we learned?

- **General methodologies & applications**
  - Guilt by association for protein function inference
  - Invariants for active site discovery
  - Emerging patterns for mutation site discovery
- **Important tactics**
  - Genome phylogenetic profiling
  - SVM-Pairwise
  - Protein-protein interactions

Any Question?

Acknowledgements

- **Some of the slides are based on slides given to me by Kenny Chua**

References

- T.F.Smith & X.Zhang. "The challenges of genome sequence annotation or `The devil is in the details'", *Nature Biotech*, 15:1222--1223, 1997
- D. Devos & A.Valencia. "Intrinsic errors in genome annotation", *TIG*, 17:429--431, 2001
- K.L.Lim et al. "Interconversion of kinetic identities of the tandem catalytic domains of receptor-like protein tyrosine phosphatase PTP-alpha by two point mutations is synergist and substrate dependent", *JBC*, 273:28986--28993, 1998
- S.F.Altshcul et al. "Basic local alignment search tool", *JMB*, 215:403--410, 1990
- S.F.Altschul et al. "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs", *NAR*, 25(17):3389--3402, 1997

References

- S.E.Brenner. "Errors in genome annotation", *TIG*, 15:132--133, 1999
- M. Pellegrini et al. "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles", *PNAS*, 96:4285--4288, 1999
- J. Wu et al. "Identification of functional links between genes using phylogenetic profiles", *Bioinformatics*, 19:1524--1530, 2003
- L.J.Jensen et al. "Prediction of human protein function from post-translational modifications and localization features", *JMB*, 319:1257--1265, 2002
- C. Wu, W. Barker. "A Family Classification Approach to Functional Annotation of Proteins", *The Practical Bioinformatician*, Chapter 19, pages 401—416, WSPC, 2004

References

- H.N. Chua, W.-K. Sung. A better gap penalty for pairwise SVM. Proc. APBC05, pages 11-20
- Hon Nian Chua, Wing Kin Sung, Limsoon Wong. Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions. *Bioinformatics*, 22:1623-1630, 2006.
- T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote homologies. *JCB*, 7(1-2):95—11, 2000
- T. Hawkins and D. Kihara. Function prediction of uncharacterized proteins. *JBCB*, 5(1):1-30, 2007