

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 8: Sequence Homology Interpretation

Limsoon Wong  
18 March 2010



2

### Plan



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

Copyright 2010 © Limsoon Wong

## Very Brief Recap of Sequence Comparison/Alignment



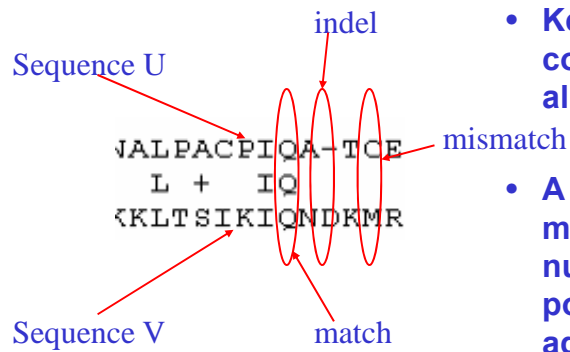
4

### 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

Copyright 2010 © Limsoon Wong

## 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**

```

      60      70      80      90      100
Amicyanin  MPHNVHFVAGVVGEEAALKGPMKKEQAYSLTFTEAGTYDYHCTFHPFMRGKVVVE
      ...: . :. :. :
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYGSLI
      70      80      90      100      110      120
  
```

No obvious match between  
Amicyanin and Ascorbate Oxidase

Copyright 2010 © Limsoon Wong



## 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 MKPGRLASIALAIIFLPMVPAHAATIEITMENLVISPTVEVSAKVGDTIRWVNKDVFAHT 60
          MK G L ++          MA PA AATIE+T++ LV SP V AKVGDTI WVN DV AHT
Sbjct: 1 MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNDVVAHT 60
  
```

good match between  
Amicyanin and unknown M. loti protein

Copyright 2010 © Limsoon Wong



## 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|      FHFTSWPDFGVFPTIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRGTGFVVIDAML
gi|2499753|     FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVHCSAGAGRTGCIYVIDIML
gi|462550|     YHHTQWPDNGVPEYALPVLTFVRRSSAARM--PETGPVIVHCSAGVGRGTGYIVIDSM
gi|2499751|     FHFTSWPDHGVPDITDILLNFRYLVRDYMKQSPPEPILVHCSAGVGRGTGTFIAIDRL
gi|1709906|     FQFTAMPDHGVPEHPTPFLAFLRRVKTGNP--PDAGPMVHCSAGVGRGTGCFIVIDAM
gi|126471|     LHFTSWPDFGVFPTIGMLKFLKKVKTLNP--VHAGPIVHCSAGVGRGTGTFIVIDAMM
gi|548626|     FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVHCSAGAGRTGCIYVIDIML
gi|131570|     FHFTGWPDHGVPYHATGLLGFVRQVKSASP--PNAGPLIVHCSAGAGRTGCFIVIDIM
gi|2144715|     FHFTSWPDHGVPDITDILLNFRYLVRDYMKQSPPEPILVHCSAGVGRGTGTFIAIDRL
          ..* *** **          *          ..***** ****... ** ..
  
```

Conserved sites

Copyright 2010 © Limsoon Wong

# Application of Sequence Comparison: Guilt-by-Association



10



## 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



Copyright 2010 © Limsoon Wong

## Function Assignment to Protein Sequence

SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR  
 YVNILPYDHSRVHLTPVEGVPSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMWE  
 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQVGD  
 VTNRKPQRLITQFHFTSWPDFGVPTPIGMLKFLKVKACNPQYAGAIVVHCSAGVGRGTG  
 TFVVIDAMLDMMHSEKVDVYGFVSRIRAQRQCMVQTDMQYVFITYQALLEHYLYGDTELE  
 VT

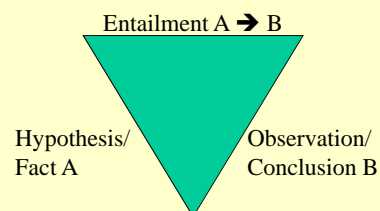
- How do we attempt to assign a function to a new protein sequence?

## 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

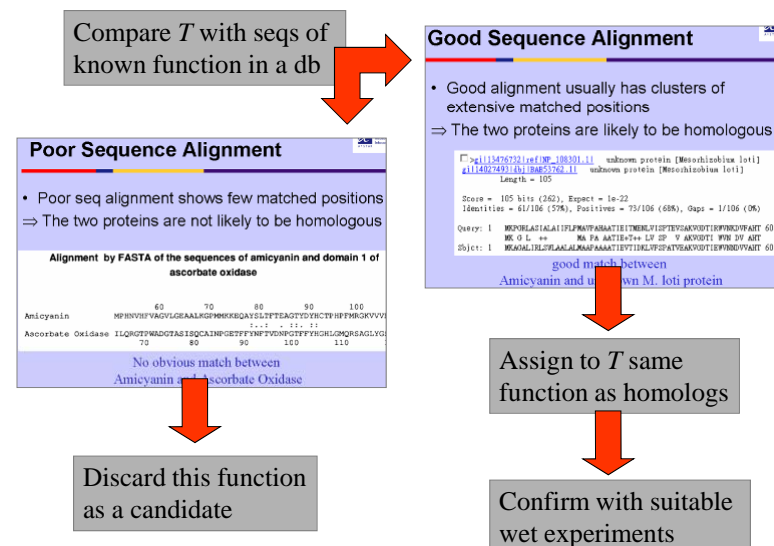


⇒ **“Guilt by association”**

## Guilt-by-Association

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

## Guilt-by-Association

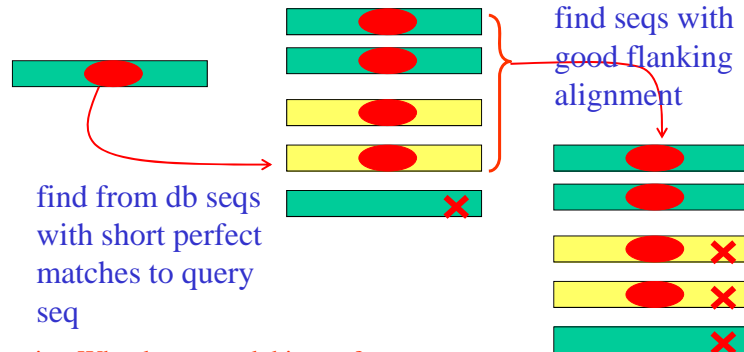


## BLAST: How It Works

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?

Copyright 2010 © Limsoon Wong

## Homologs obtained by BLAST



Sequences producing significant alignments:	Score (bits)	E Value
<a href="#">gi 14193729 gb AAK56109.1 AF332081.1</a> protein tyrosin phosph...	62.1	e-177
<a href="#">gi 126467 sp P18433 PTRA_HUMAN</a> Protein-tyrosine phosphatase...	62.1	e-177
<a href="#">gi 4506303 ref NP_002827.1 </a> protein tyrosine phosphatase, r...	62.1	e-176
<a href="#">gi 227294 prf 11701300A</a> protein Tyr phosphatase	62.0	e-176
<a href="#">gi 18450369 ref NP_543030.1 </a> protein tyrosine phosphatase, ...	62.1	e-176
<a href="#">gi 32067 emb CAA37447.1 </a> tyrosine phosphatase precursor [Ho...	61.1	e-176
<a href="#">gi 285113 pir JC1285</a> protein-tyrosine-phosphatase (EC 3.1....	61.9	e-176
<a href="#">gi 6981446 ref NP_036895.1 </a> protein tyrosine phosphatase, r...	61.1	e-176
<a href="#">gi 2098414 pdb 1YFO A</a> Chain A, Receptor Protein Tyrosine Ph...	61.1	e-174
<a href="#">gi 32313 emb CAA38662.1 </a> protein-tyrosine phosphatase [Homo...	61.1	e-174
<a href="#">gi 450583 gb AA04150.1 </a> protein tyrosine phosphatase >gi 4...	60.5	e-172
<a href="#">gi 6679557 ref NP_033006.1 </a> protein tyrosine phosphatase, r...	60.1	e-172
<a href="#">gi 483922 gb AAA17990.1 </a> protein tyrosine phosphatase alpha	59.9	e-170

- Thus our example sequence could be a protein tyrosine phosphatase  $\alpha$  (PTP $\alpha$ )

Copyright 2010 © Limsoon Wong



## Example Alignment with PTP $\alpha$



Score = 632 bits (1629), Expect = 0.180  
 Identities = 294/302 (97%), Positives = 294/302 (97%)

```

Query: 1  SPSTNRKYPPFLFVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASXXXXXXXXX 61
          SPSTNRKYPPFLFVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAAS      R
Sbjct: 232 SPSTNRKYPPFLFVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR 261

Query: 61  YVNILPYDHSRVHLTPVEGVPSDYINASFINGYQEKKNFIAAQGPKEETVNDFWRMIVE 120
          YVNILPYDHSRVHLTPVEGVPSDYINASFINGYQEKKNFIAAQGPKEETVNDFWRMIVE
Sbjct: 252 YVNILPYDHSRVHLTPVEGVPSDYINASFINGYQEKKNFIAAQGPKEETVNDFWRMIVE 321

Query: 121 QNTAIVMVINLKERKECKCAQYWPDQGCWYGNVRSVSDVTVLVDYTVRKFCIQQVGD 180
          QNTAIVMVINLKERKECKCAQYWPDQGCWYGNVRSVSDVTVLVDYTVRKFCIQQVGD
Sbjct: 322 QNTAIVMVINLKERKECKCAQYWPDQGCWYGNVRSVSDVTVLVDYTVRKFCIQQVGD 381

Query: 131 VTNREPQLITCFHFTSWPDFGVFFTPIGMLKFLKVKACNPQYAGAI7VHCSAGVGRGTG 240
          VTNREPQLITCFHFTSWPDFGVFFTPIGMLKFLKVKACNPQYAGAI7VHCSAGVGRGTG
Sbjct: 332 VTNREPQLITCFHFTSWPDFGVFFTPIGMLKFLKVKACNPQYAGAI7VHCSAGVGRGTG 441

Query: 241 TFVV.DAMLDMMHSERKVDVYGFVSRIRAQRCQIVQTDMQVVFVYQALLEHYLYGDTLE 300
          TFVV.DAMLDMMHSERKVDVYGFVSRIRAQRCQIVQTDMQVVFVYQALLEHYLYGDTLE
Sbjct: 442 TFVV.DAMLDMMHSERKVDVYGFVSRIRAQRCQIVQTDMQVVFVYQALLEHYLYGDTLE 501
  
```

Copyright 2010 © Limsoon Wong

## 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

Copyright 2010 © Limsoon Wong



## 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%



## 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

## Lightning Does Strike Twice!

- **Roy Sullivan, a former park ranger from Virginia, 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**



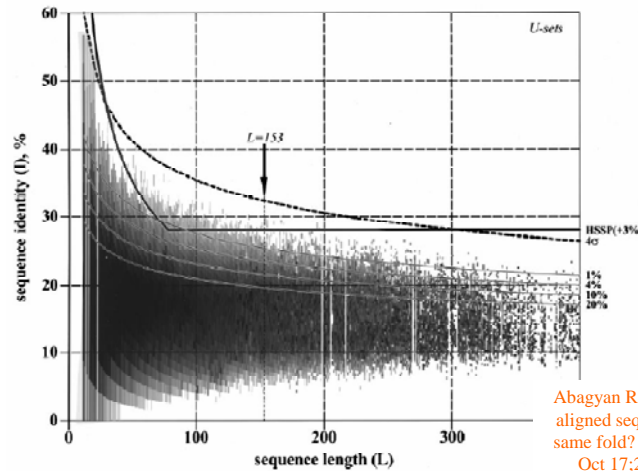
Cartoon: Ron Hipschman  
Data: David Hand

## 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**



## 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

Copyright 2010 © Limsoon Wong



## Examples of Invalid Function Assignment: The IMP Dehydrogenases (IMPDH)

18 entries were found

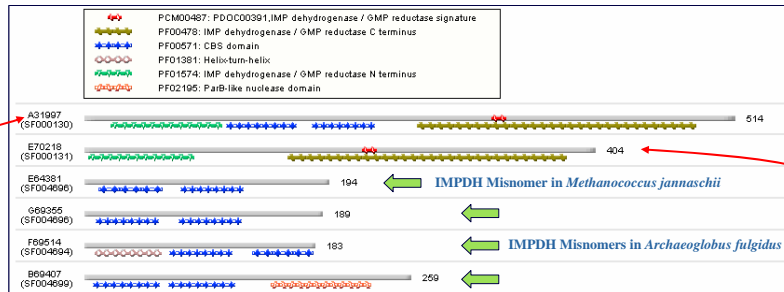
ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept
NF00181837	Methanococcus jannaschii	<a href="#">E64831</a> conserved hypothetical protein MJ0653	<a href="#">Y653_MET1A</a> Hypothetical protein MJ0653	<a href="#">g1_32230</a> inosine 3'-monophosphate dehydrogenase (guaB) <a href="#">NP_247637</a> inosine 3'-monophosphate dehydrogenase (guaB)
NF00187788	Archaeoglobus fulgidus	<a href="#">C69355</a> MJ0653 homolog AF0847 <a href="#">ALT_NAM652</a> inosine monophosphate dehydrogenase (guaB-1) homolog [misnomer]	<a href="#">C26411</a> INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)	<a href="#">g2649754</a> inosine monophosphate dehydrogenase (guaB-1) <a href="#">NP_089261</a> inosine monophosphate dehydrogenase (guaB-1)
NF00188267	Archaeoglobus fulgidus	<a href="#">E69514</a> yhcV homolog 2 <a href="#">ALT_NAM653</a> inosine monophosphate dehydrogenase (guaB-2) homolog [misnomer]	<a href="#">C28162</a> INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-2)	<a href="#">g26498410</a> inosine monophosphate dehydrogenase (guaB-2) <a href="#">NP_070943</a> inosine monophosphate dehydrogenase (guaB-2)
NF00188697	Archaeo			inosophosphate ve nophosphate pe
NF00197776	Thermo			inosophosphate d protein nophosphate d protein
NF00414709	Methanothermobacter thermoautotrophicus	<a href="#">E69933</a> <a href="#">E69933</a> inosine monophosphate dehydrogenase related protein V [misnomer]	<a href="#">C27284</a> INOSINE 3'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	dehydrogenase related protein V <a href="#">NP_276315</a> inosine 3'-monophosphate dehydrogenase related protein V
NF00414811	Methanothermobacter thermoautotrophicus	<a href="#">E69933</a> M11232 protein homolog MTH126 <a href="#">ALT_NAM653</a> inosine-3'-monophosphate dehydrogenase related protein VII [misnomer]	<a href="#">C26720</a> INOSINE-3'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII	<a href="#">g2621166</a> inosine-3'-monophosphate dehydrogenase related protein VII <a href="#">NP_275240</a> inosine-3'-monophosphate dehydrogenase related protein VII
NF00414837	Methanothermobacter thermoautotrophicus	<a href="#">E69232</a> M11225-related protein MTH1992 <a href="#">ALT_NAM653</a> inosine-3'-monophosphate dehydrogenase related protein IX [misnomer]	<a href="#">C27073</a> INOSINE-3'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN IX	<a href="#">g2622023</a> inosine-3'-monophosphate dehydrogenase related protein IX <a href="#">NP_276137</a> inosine-3'-monophosphate dehydrogenase related protein IX
NF00414969	Methanothermobacter thermoautotrophicus	<a href="#">E69277</a> yhcV homolog 2 <a href="#">ALT_NAM653</a> inosine monophosphate dehydrogenase related protein X [misnomer]	<a href="#">C27616</a> INOSINE-3'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN X	<a href="#">g2622697</a> inosine-3'-monophosphate dehydrogenase related protein X <a href="#">NP_276687</a> inosine-3'-monophosphate dehydrogenase related protein X

**A partial list of IMP dehydrogenase misnomers in complete genomes remaining in some public databases**

Copyright 2010 © Limsoon Wong



# 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



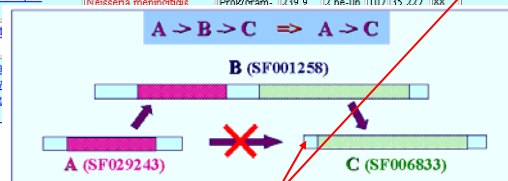
# Invalid Transitive Assignment

Root of invalid transitive assignment

Accession	Gene ID	Protein Name	Organism	EC	Length	Similarity	Assignment
H70468	SF001258	051440	phosphonobosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphonobosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Aquifex aeolicus	594.3	4.8e-26	205 39,086 197
S76963	SF001258	039935	phosphonobosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphonobosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Synechocystis sp.	557.0	5.7e-24	230 39,175 194
T35073	SF029243	005738	probable phosphonobosyl-AMP cyclohydrolase	Streptomyces coelicolor	399.3	3.5e-15	128 42,157 102
S53349	SF001257	001188	phosphonobosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphonobosyl-ATP pyrophosphatase (EC 3.6.1.31) / histidinol dehydrogenase (EC 1.1.1.23)	Saccharomyces cerevisiae	384.1	2.5e-14	799 31,863 204
E69493	SF029243	005738	phosphonobosyl-AMP cyclohydrolase (EC 3.5.4.19) [similarity]	Archaeoglobus fulgidus	396.8	4.8e-15	108 47,778 90
G64337	SF006833	030827	phosphonobosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Methanococcus jannaschii	246.9	1.1e-06	95 36,842 95
D81178	SF006833	101491	phosphonobosyl-ATP pyrophosphatase (EC 3.6.1.31) NMB0603 [similarity]	Neisseria meningitidis	239.9	2.6e-06	107 35,227 88
G81925	SF006833	101491	phosphonobosyl-ATP pyrophosphatase (EC 3.6.1.31) NMA0807 [similarity]				
S51513	SF001257	001188	phosphonobosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphonobosyl-ATP pyrophosphatase (EC 3.6.1.31) / histidinol dehydrogenase (EC 1.1.1.23)				

B →  
A →  
C →

Mis-assignment of function

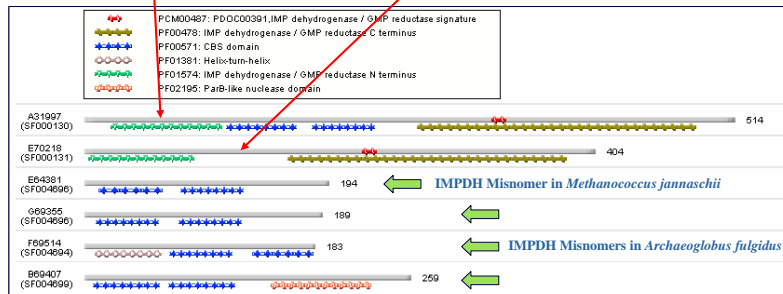


No IMPDH domain

## Emerging Pattern

Typical IMPDH

Functional IMPDH w/o CBS



- Most IMPDHs have 2 IMPDH and 2 CBS domains
  - Some IMPDH (E70218) lacks CBS domains
- ⇒ IMPDH domain is the emerging pattern

Copyright 2010 © Limsoon Wong

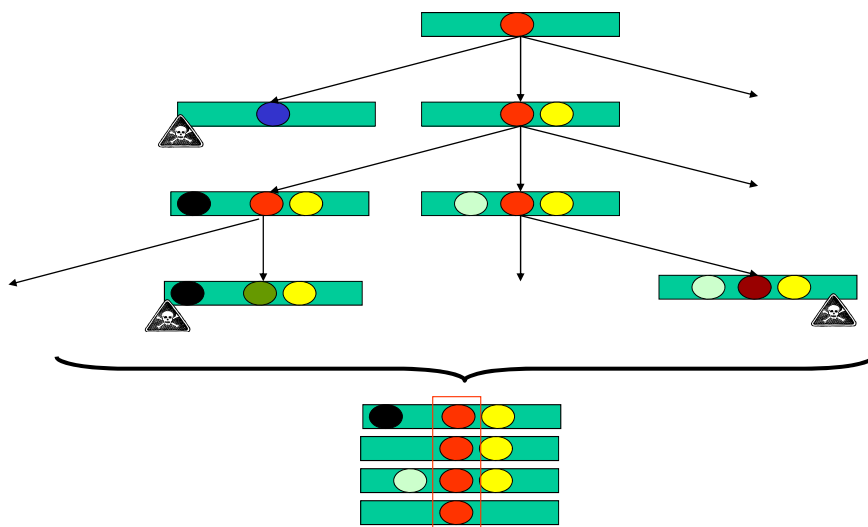
Application of  
Sequence Comparison:  
Active Site/Domain Discovery

## 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?

## In the course of evolution...



## Multiple Alignment of PTPs

```

gi|126467|      FHFTSVPDFGVPFTP I GMLKFLKKVKACNP--QYAGAIVVHCSAGVGRGTGFVVIDAMLD
gi|2499753|    FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|462550|     YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRGTGTYIVIDSMLQ
gi|2499751|    FHFTSVPDHGVPD TDDLINFRYLVRD YMKQSPPEP I LVHCSAGVGRGTGTF IAIDRLIY
gi|1709906|    FQFTA WPDHGVP EHP T PFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRGTGCF IVIDAMLE
gi|126471|     LHFTSVPDFGVPFTP I GMLKFLKKVKTLNMP--VHAGPIVVHCSAGVGRGTGTF IVIDAMMA
gi|548626|     FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|131570|     FHFTGWPDHGVPYHATGLLGFVRQVKS KSP--PNAGPLVVHCSAGAGRTGCF IVIDIMLD
gi|2144715|    FHFTSVPDHGVPD TDDLINFRYLVRD YMKQSPPEP I LVHCSAGVGRGTGTF IAIDRLIY
..* *** **          . *                ..***** ***** ** ..

```

- Notice the PTPs agree with each other on some positions more than other positions
  - These positions are more imp't 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?

genome phylogenetic profiles  
protfun's feature profiles  
Similarity of dissimilarities



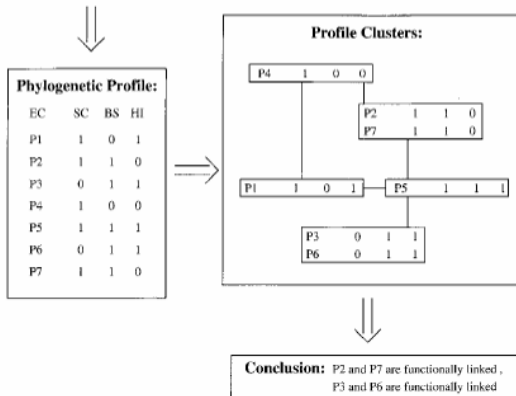
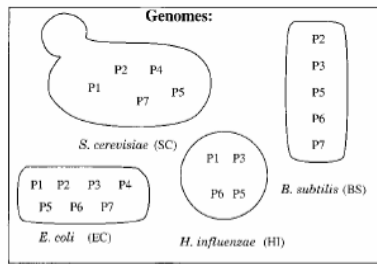
## What if there is no useful seq homolog?

- **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

## 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**



## 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

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

where

No. of ways to distribute  $z$  co-occurrences over  $N$  lineage's  $w_z = \binom{N}{z}$

No. of ways to distribute the remaining  $x - z$  and  $y - z$  occurrences over the remaining  $N - z$  lineage's  $\bar{w}_z = \binom{N - z}{x - z} * \binom{N - z}{y - z}$

No. of ways of distributing  $X$  and  $Y$  over  $N$  lineage's without restriction  $W = \binom{N}{x} * \binom{N}{y}$

## Phylogenetic Profiles: Evidence

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



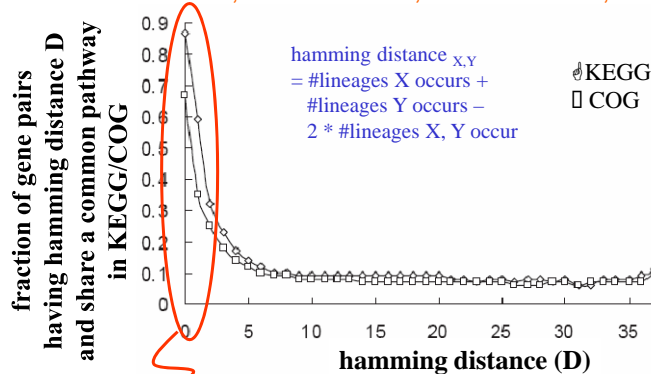
Keyword	No. of non-homologous proteins in group	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins <sup>†</sup>	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdopterin and Molybdenum, and molybdopterin	12	6	1
Hypothetical <sup>‡</sup>	1,084	198,226	8,440

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

Copyright 2010 © Limsoon Wong

## 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?

Copyright 2010 © Limsoon Wong





## Guilt by Association of Dissimilarities



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



"unknown" is an "apple"!

	Orange <sub>1</sub>	Banana <sub>1</sub>	...
Apple <sub>1</sub>	 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 <sub>2</sub>	 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 <sub>1</sub>	 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

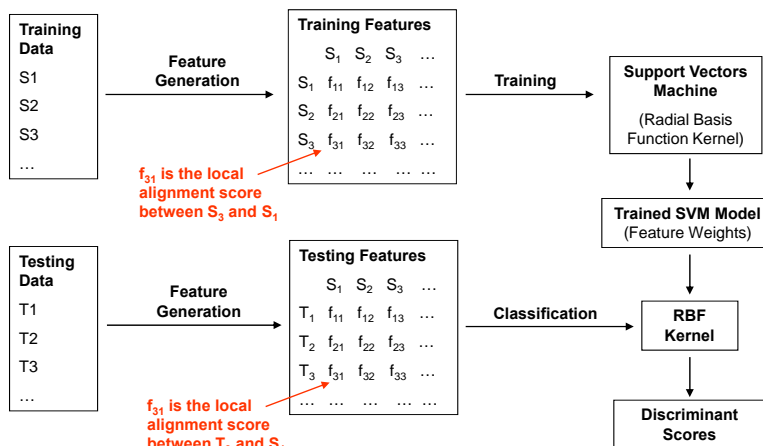
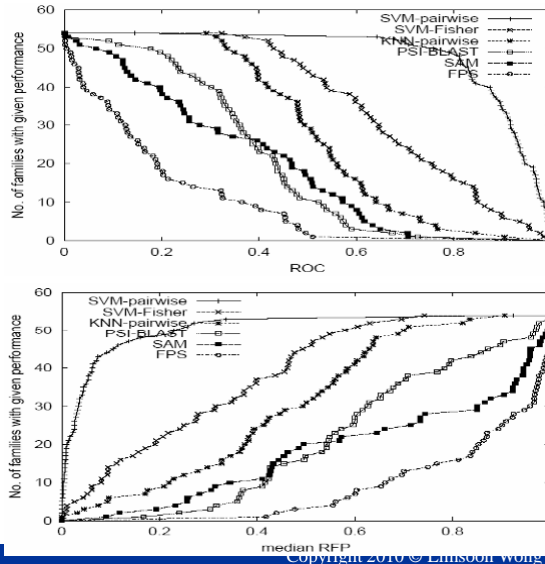


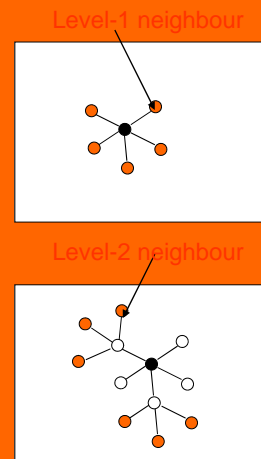
Image credit: Kenny Chua

## 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.

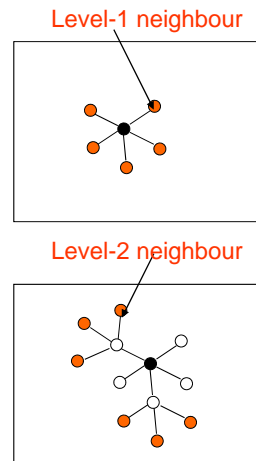


## Protein Function Prediction from Protein Interactions



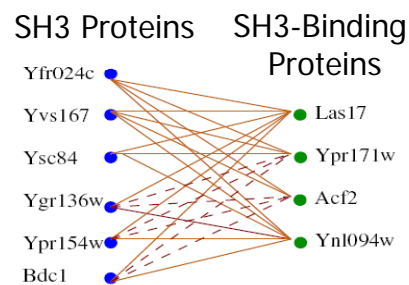
## Functional Association Thru Interactions

- **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



Copyright 2010 © Limsoon Wong

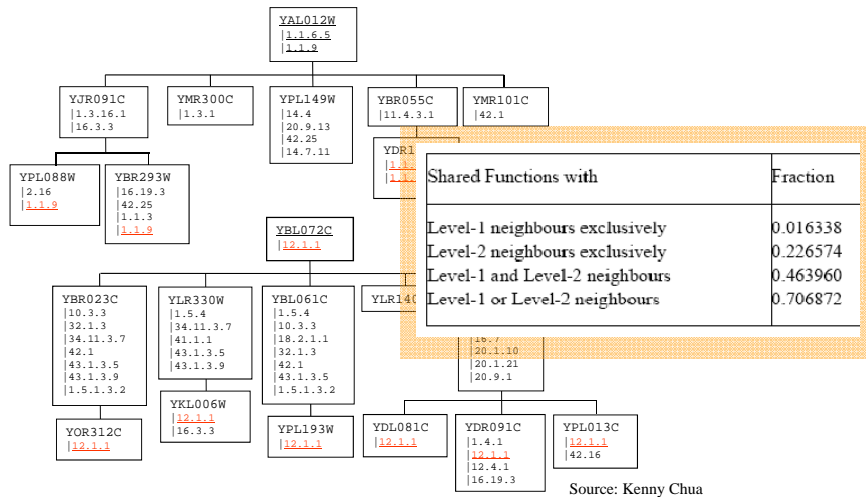
## An illustrative Case of Indirect Functional Association?



- 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?

Copyright 2010 © Limsoon Wong

## Freq of Indirect Functional Association

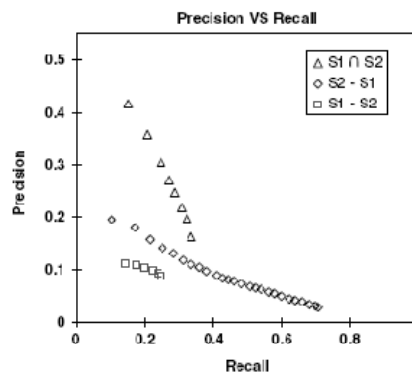


## Prediction Power By Majority Voting

- 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

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

- $n_i$  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  $k$
- $X \Delta Y$  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**

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

Copyright 2010 © Limsoon Wong

## 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**

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

Copyright 2010 © Limsoon Wong



## Correlation w/ Functional Similarity

- Correlation betw functional similarity & estimates

Neighbours	CD-Distance	FS-Weight
S <sub>1</sub>	0.471810	0.498745
S <sub>2</sub>	0.224705	0.298843
S <sub>1</sub> ∪ S <sub>2</sub>	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<sub>i</sub> is reliability of expt source i,
- E<sub>u,v</sub> 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 between  $u$  and  $v$

⇒ Rewriting

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

Copyright 2010 © Limsoon Wong

## Integrating Reliability

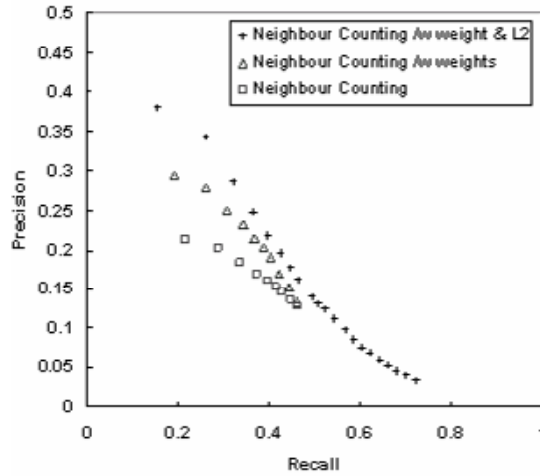


- Equiv measure shows improved correlation w/ functional similarity when reliability of interactions is considered:

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

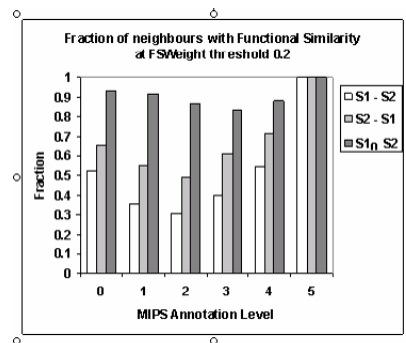
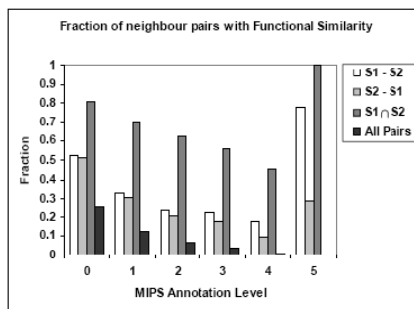
Copyright 2010 © Limsoon Wong

## Improvement to Prediction Power by Majority Voting



Considering only neighbours w/ FS weight > 0.2

## Improvement to Over-Rep of Functions in Neighbours



## Use L1 & L2 Neighbours for Prediction

- **FS-weighted Average**

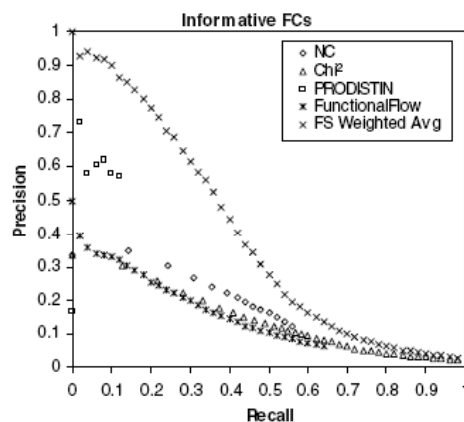
$$f_x(u) = \frac{1}{Z} \left[ \lambda r_{int} \pi_x + \sum_{v \in N_u} \left( S_{TR}(u, v) \delta(v, x) + \sum_{w \in N_v} S_{TR}(u, w) \delta(w, x) \right) \right]$$

- $r_{int}$  is fraction of all interaction pairs sharing function
- $\lambda$  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  $k$
- $\pi_x$  is freq of function  $x$  in the dataset
- $Z$  is sum of all weights

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

## Performance of FS-Weighted Averaging

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



# Application of Sequence Comparison: Key Mutation Site Discovery



59

## Identifying Key Mutation Sites



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

Sequence from a typical PTP domain D2

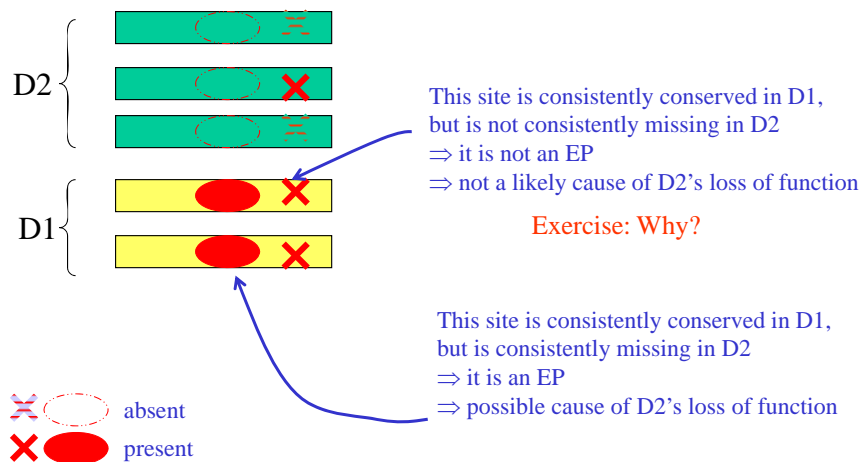
```
>gi|00000|PTP1-D2  
EEEFKKLTSIKIQNDKERTGNLPANHEKKNVQLQIIPYEFNEVIIPVKEGZLNTDYNASF  
IDGYRQKDSYILASQGPLLETIEDFWRHIEWRKSCSIVELTELEERGOERC&QYWPSDGLV  
SYGDITVELKKEECCESYTVRDLVINTRENKSEQIQPFHFGWPEVVGIPSDGKGLSII  
AIVQKQQQSGNHPITVHCSSAGCRTGTFCALSTVLERVKLEGILDVYQTVKSLRLQRP  
HWQTLQYEFPCYKVVQEVYIDAFSDYANFK
```

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

## 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

## Emerging Patterns of PTP D1 vs D2



## Key Mutation Site: PTP D1 vs D2



```

? ! ? ? ? ? ? ?
gi|00000|P D2 QFHFGWPEVGI PSDGKGMISIIAAVQKQQQ--SGNHPITVHCSAGAGRTGTF CALSTVL
gi|126467| QFHFTSWP DFGVPF TPIGMLKFLKKVKACNP--QYAGAI VVHCSAGVGR TGFVVIDAML
gi|2499753 QFHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGP I VVHCSAGAGRTG C Y IVID IML
gi|462550| QYHYTQWPD MGVP EYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGR TGTY IVID SML
gi|2499751 QFHFTSWPDHGVPD TTDLL INFRYLVRD YMKQSPPE SPILVHCSAGVGR TGTF I AIDRL I
gi|1709906 D1 QFQFTA WPDHGVP EHP T PFLAFLRRVKT CNP--PDAGPMVVHCSAGVGR TGCF IVID AML
gi|126471| QLHFTSWP DFGVPF TPIGMLKFLKKVKT LNP--VHAGP I VVHCSAGVGR TGTF IVID AMM
gi|548626| QFHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGP I VVHCSAGAGRTG C Y IVID IML
gi|131570| QFHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLV VHCSAGAGRTG C F IVID IML
gi|2144715 QFHFTSWPDHGVPD TTDLL INFRYLVRD YMKQSPPE SPILVHCSAGVGR TGTF I AIDRL I
* .. ** . *. * . . ***** ** .. . .

```

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

Copyright 2010 © Limsoon Wong

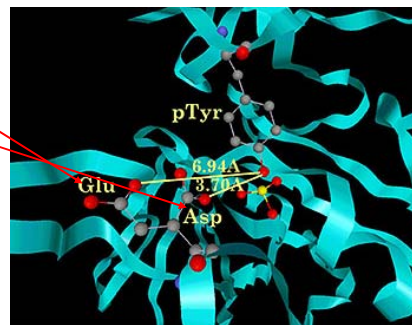
## Key Mutation Site: PTP D1 vs D2



```

? ! ?
gi|00000|P D2 QFHFGWPEVGI PSDGK
gi|126467| QFHFTSWP DFGVPF TPI
gi|2499753 QFHFTGWPDHGVPYHAT
gi|462550| QYHYTQWPD MGVP EYAL
gi|2499751 QFHFTSWPDHGVPD TTD
gi|1709906 D1 QFQFTA WPDHGVP EHP T
gi|126471| QLHFTSWP DFGVPF TPI
gi|548626| QFHFTGWPDHGVPYHAT
gi|131570| QFHFTGWPDHGVPYHAT
gi|2144715 QFHFTSWPDHGVPD TTD
* .. ** . *. *

```



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

Copyright 2010 © Limsoon Wong

## Confirmation by Mutagenesis Expt



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

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

## 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 Questions?

## 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.Altshcul 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