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**

Copyright 2010 © Limsoon Wong

Very Brief Recap of Sequence Comparison/Alignment

Motivations for Sequence Comparison

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

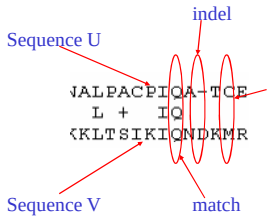
⇒ **Evolution is related to changes in DNA**

⇒ **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**

Copyright 2010 © Limsoon Wong

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**

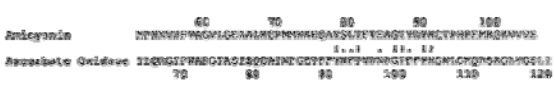
Copyright 2010 © Limsoon Wong

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



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|134767321|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   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 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

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|      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

Copyright 2010 © Limsoon Wong

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**

Copyright 2010 © Limsoon Wong

Function Assignment to Protein Sequence

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

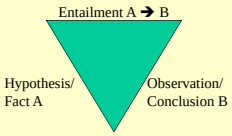
- **How do we attempt to assign a function to a new protein sequence?**

Copyright 2010 © Limsoon Wong

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"**

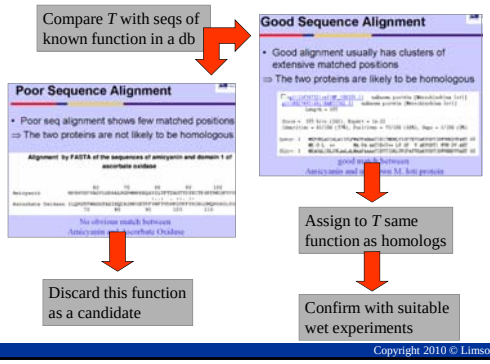
Copyright 2010 © Limsoon Wong

Guilt-by-Association

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

Copyright 2010 © Limsoon Wong

Guilt-by-Association

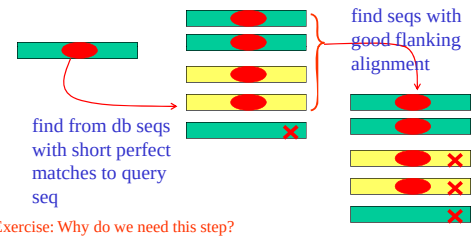


Copyright 2010 © Limsoon Wong

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    E
                                                                     (bits) Value

gi|14193729|gb|AAK56109.1|AF332081_1  protein tyrosin phosph...      62 L   e-177
gi|126467|sp|P18433|PTPA_HUMAN  Protein-tyrosine phosphatase...      62 L   e-177
gi|4506303|ref|NP_002827.1|  protein tyrosine phosphatase, r...      62 L   e-176
gi|2272794|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|285115|pir||JC1285  protein-tyrosine-phosphatase (EC 3.1....      619    e-176
gi|16981446|ref|NP_036895.1|  protein tyrosine phosphatase, r...     61 L   e-176
gi|20984141|pdb|1YFO|A  Chain A, Receptor Protein Tyrosine Ph...     61 S   e-174
gi|32313|emb|CAA30662.1|  protein-tyrosine phosphatase [Homo...      61 L   e-174
gi|14505831|gb|AAB04150.1|  protein tyrosine phosphatase >gi|4...    605    e-172
gi|16679557|ref|NP_033006.1|  protein tyrosine phosphatase, r...     60 L   e-172
gi|4839221|gb|AAA17990.1|  protein tyrosine phosphatase alpha        599    e-170
```

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

Copyright 2010 © Limsoon Wong

Example Alignment with PTPα

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

Query: 1    SPSTNRKVPPLPVDKLEEEINKHMADDNKLFREEFNALPACPIQATCEAAKKLGSSDLDCKR 60
            SPSTNRKVPPLPVDKLEEEINKHMADDNKLFREEFNALPACPIQATCEAAK                R
Sbjct: 202  SPSTNRKVPPLPVDKLEEEINKHMADDNKLFREEFNALPACPIQATCEAASKEEEKKKKKK 261

Query: 61   YVNILPYDHSRVHLTPVEGVPDSDYINASPIMQYQEKNKFIAAQGPKEETVNDFWRMIWE 120
            YVNILPYDHSRVHLTPVEGVPDSDYINASPIMQYQEKNKFIAAQGPKEETVNDFWRMIWE
Sbjct: 262  YVNILPYDHSRVHLTPVEGVPDSDYINASPIMQYQEKNKFIAAQGPKEETVNDFWRMIWE 321

Query: 121  QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 180
            QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
Sbjct: 322  QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 381

Query: 181  VTNSKFQSLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG 240
            VTNSKFQSLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
Sbjct: 382  VTNSKFQSLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG 441

Query: 241  TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLHHYLFGDTSLE 300
            TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLHHYLFGDTSLE
Sbjct: 442  TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLHHYLFGDTSLE 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%

Copyright 2010 © Limsoon Wong

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

Copyright 2010 © Limsoon Wong

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

Copyright 2010 © Limsoon Wong

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

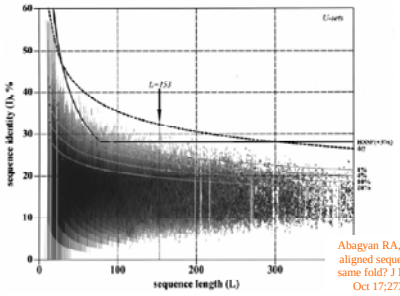
⇒ While it is worth noting that two proteins contain similar low complexity regions, they are best excluded when constructing alignments

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

Source: NCBI

Copyright 2010 © Limsoon Wong

Effect of Sequence Length



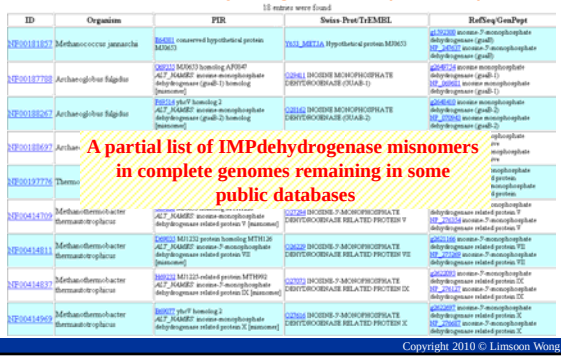
Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol. 1997 Oct 17;273(1):355-68

Copyright 2010 © Limsoon Wong

Examples of Invalid Function Assignment:

The IMP Dehydrogenases (IMPDH)

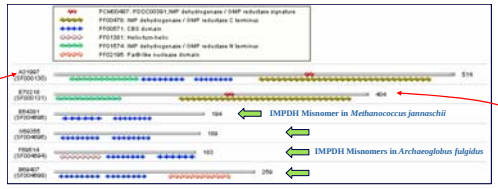
A partial list of IMPdehydrogenase misnomers in complete genomes remaining in some public databases



ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept
NF00181857	Methanococcus jannaschii	B64381 conserved hypothetical protein M30653	Q57_METJA Hypothetical protein M30653	gi|39288 inosine-5'-monophosphate dehydrogenase (gsdI) NP_247627 inosine-5'-monophosphate dehydrogenase (gsdI)
NF00187788	Archaeoglobus fulgidus	Q48231 M30653 homolog AF0347 ALT_NAMEF inosine-monophosphate dehydrogenase (gsdI-1) homolog [inosicase]	O28841 INOSINE MONOPHOSPHATE DEHYDROGENASE (IUAB-1)	gi|2648724 inosine monophosphate dehydrogenase (gsdI-1) NP_069681 inosine monophosphate dehydrogenase (gsdI-1)
NF00188267	Archaeoglobus fulgidus	P48214 yhcF homolog 2 ALT_NAMEF inosine-monophosphate dehydrogenase (gsdI-2) homolog [inosicase]	O28142 INOSINE MONOPHOSPHATE DEHYDROGENASE (IUAB-2)	gi|2648402 inosine monophosphate dehydrogenase (gsdI-2) NP_070843 inosine monophosphate dehydrogenase (gsdI-2)
NF00188697	Archae-			ophosphate ase monophosphate ase
NF00197776	Thermo			monophosphate d protein monophosphate d protein
NF00414709	Methanothermobacter thermautotrophicus	ALT_NAMEF inosine-monophosphate dehydrogenase related protein V [inosicase]	O27284 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	monophosphate dehydrogenase related protein V NP_276354 inosine-5'-monophosphate dehydrogenase related protein V
NF00414811	Methanothermobacter thermautotrophicus	D69021 M11233 protein homolog MTH126 ALT_NAMEF inosine-5'-monophosphate dehydrogenase related protein VII [inosicase]	O28229 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII	gi|2621166 inosine-5'-monophosphate dehydrogenase related protein VII NP_277269 inosine-5'-monophosphate dehydrogenase related protein VII
NF00414837	Methanothermobacter thermautotrophicus	H69231 M11233-related protein MTH992 ALT_NAMEF inosine-5'-monophosphate dehydrogenase related protein IX [inosicase]	O27072 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN IX	gi|2622092 inosine-5'-monophosphate dehydrogenase related protein IX NP_276127 inosine-5'-monophosphate dehydrogenase related protein IX
NF00414969	Methanothermobacter thermautotrophicus	B69077 yhcF homolog 2 ALT_NAMEF inosine-monophosphate dehydrogenase related protein X [inosicase]	O27616 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN X	gi|2622097 inosine-5'-monophosphate dehydrogenase related protein X NP_276627 inosine-5'-monophosphate dehydrogenase related protein X

Copyright 2010 © Limsoon Wong

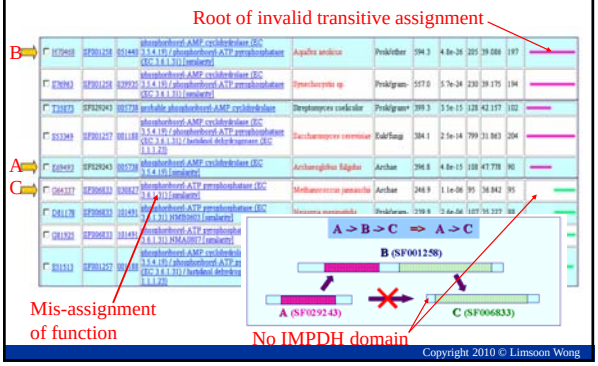
IMPDH Domain Structure



- **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**

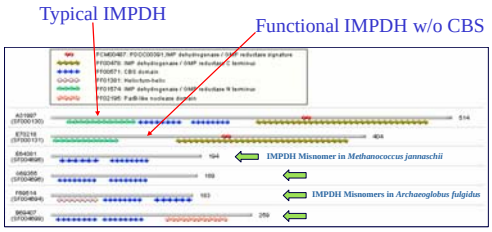
Copyright 2010 © Limsoon Wong

Invalid Transitive Assignment



Copyright 2010 © Limsoon Wong

Emerging Pattern



- **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

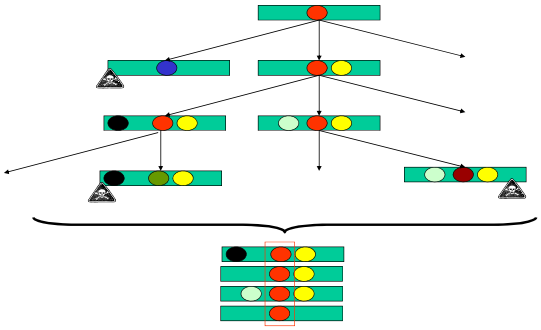
Discover Active Site and/or Domain

- **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?

Copyright 2010 © Limsoon Wong

In the course of evolution…



Copyright 2010 © Limsoon Wong

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**

Copyright 2010 © Limsoon Wong

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

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

Copyright 2010 © Limsoon Wong

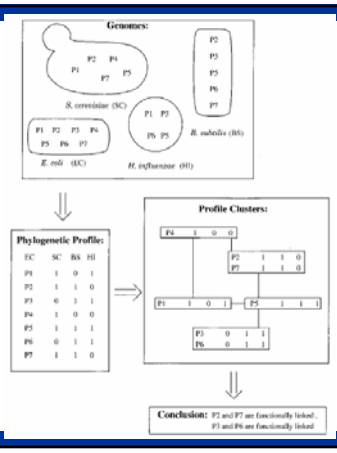
Phylogenetic Profiling

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**

Copyright 2010 © Limsoon Wong



Phylogenetic Profiling: How it Works

Copyright 2010 © Limsoon Wong

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

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

where

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

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

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

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

Copyright 2010 © Limsoon Wong

Phylogenetic Profiles: Evidence

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

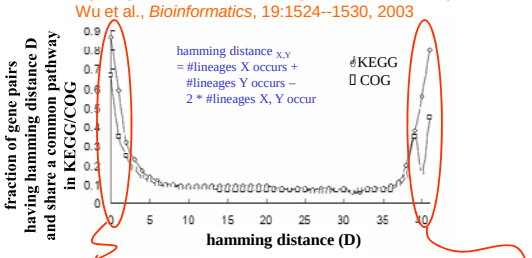
Keyword	No. of non-homologous proteins in group	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin and Molybdenum, and molybdeterin	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**

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$_1$	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**	...
...	...	...	...

Copyright 2010 © Limsoon Wong

SVM-Pairwise Framework

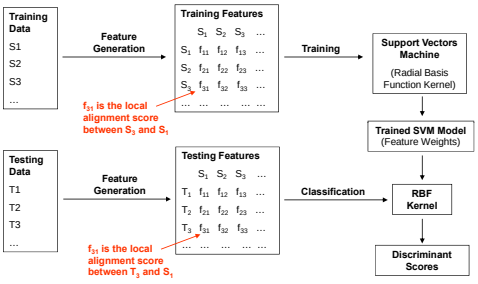
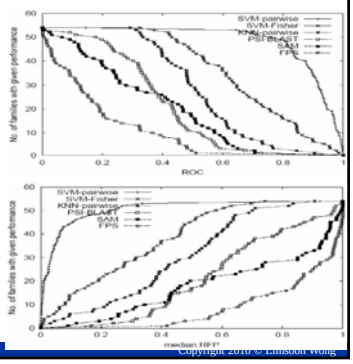


Image credit: Kenny Chua

Copyright 2010 © Limsoon Wong

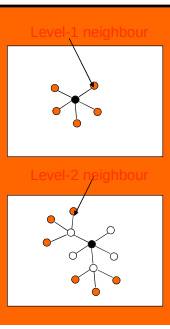
Performance of SVM-Pairwise

- **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.



Copyright 2010 © Limsoon Wong

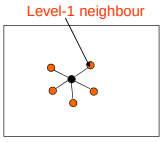
Protein Function Prediction from Protein Interactions



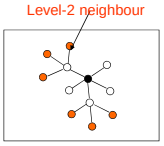
Functional Association Thru Interactions

- **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

Level-1 neighbour

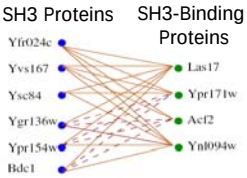


Level-2 neighbour



Copyright 2010 © Limsoon Wong

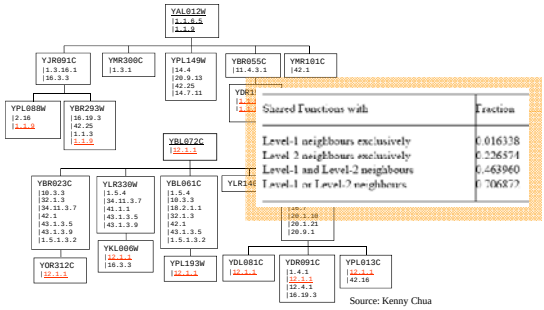
An illustrative Case of Indirect Functional Association?



- **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



Source: Kenny Chua

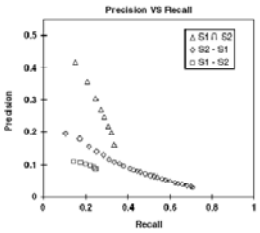
Copyright 2010 © Limsoon Wong

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**

Copyright 2010 © Limsoon Wong

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**

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

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_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**

Copyright 2010 © Limsoon Wong

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**

Copyright 2010 © Limsoon Wong

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**

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

Copyright 2010 © Limsoon Wong

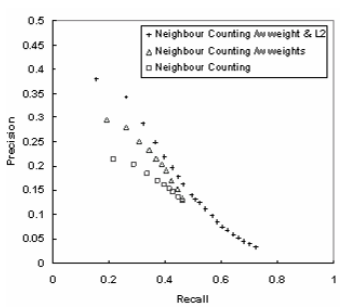
Integrating Reliability

- **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

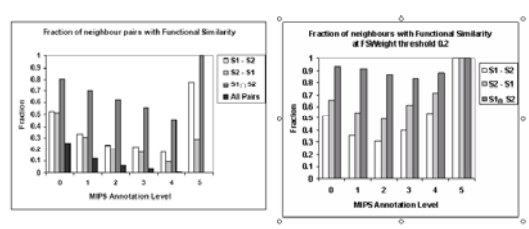
Improvement to Prediction Power by Majority Voting



Considering only neighbours w/ FS weight > 0.2

Copyright 2010 © Limsoon Wong

Improvement to Over-Rep of Functions in Neighbours



Copyright 2010 © Limsoon Wong

Use L1 & L2 Neighbours for Prediction

- **FS-weighted Average**

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

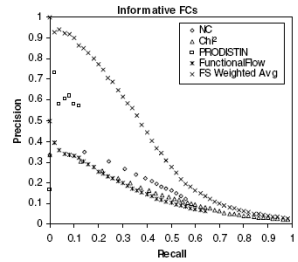
  - r_{int} is fraction of all interaction pairs sharing function
  - λ is weight of contribution of background freq
  - $\delta(k, x) = 1$ if k has function x, 0 otherwise
  - N_k is the set of interacting partners of k
  - π_x is freq of function x in the dataset
  - Z is sum of all weights

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

Copyright 2010 © Limsoon Wong

Performance of FS-Weighted Averaging

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



Copyright 2010 © Limsoon Wong

About the Inventor: Chua Hon Nian

- **Chua Hon Nian**
  - PhD, NUS, 2008
  - Currently postdoc at Harvard
  - 49th hottest paper in Computer Science published in 2006
  - Winner, DREAM2 challenge PPI subnetwork, 2007

Copyright 2010 © Limsoon Wong

Application of Sequence Comparison: Key Mutation Site Discovery

Identifying Key Mutation Sites

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

Sequence from a typical PTP domain D2

```
>gi|00000|PTPA-D2
ZZZYKKLTSIKIQNDKERTGNLPANEKENEVLQIIPYEFNEVIIPVERGEENTDYFNASF
IDGYRQKDSYIASQGPLLETIEDFWRMIWEWKSCSIVMLTELERGQEKCAQYWPSDGLV
SYGDITVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFHGWPEVGIPSDGKGMISII
AAVQKQQQQSGNHPITVHCSAGAGRTGTFCALSTVLERVKAEGILDVFQTVKSLRLQRPH
HVQTLEQYEFCYKVVQEYIDAFSDYANYK
```

- **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?**

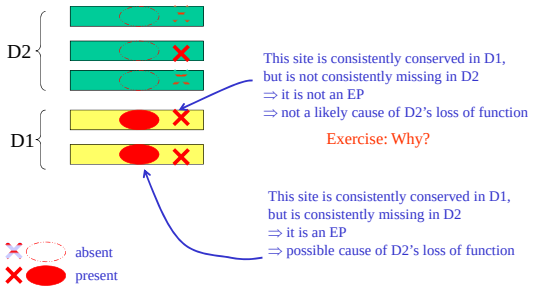
Copyright 2010 © Limsoon Wong

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**

Copyright 2010 © Limsoon Wong

Emerging Patterns of PTP D1 vs D2



Copyright 2010 © Limsoon Wong

Key Mutation Site: PTP D1 vs D2

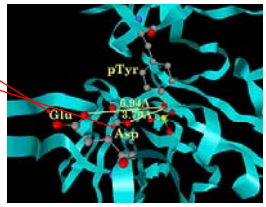
```
                 ?   !  ?          ?                    ?              ? ??
gi|00000|P  D2  QFHFHGWPEVGIPSDGKGMISIIAAVQKQQQQ-SGNHPITVHCSAGAGRTGTFCALSTVL
gi|126467|      QFHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAML
gi|2499753      QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML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

Copyright 2010 © Limsoon Wong

Key Mutation Site: PTP D1 vs D2

```
                 ?   !  ?
gi|00000|P  D2  QFHFHGWPEVGIPSDGK
gi|126467|      QFHFTSWPDFGVPFTPI
gi|2499753      QFHFTGWPDHGVPYHAT
gi|462550|      QYHYTQWPDMGVPEYAL
gi|2499751      QFHFTSWPDHGVPDTTD
gi|1709906  D1  QFQFTAWPDHGVPEHPT
gi|126471|      QLHFTSWPDFGVPFTPI
gi|548626|      QFHFTGWPDHGVPYHAT
gi|131570|      QFHFTGWPDHGVPYHAT
gi|2144715      QFHFTSWPDHGVPDTTD
                * ..  **. *.*
```



- **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 → 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?

Copyright 2010 © Limsoon Wong

About the Inventor: Prasanna Kolatkar

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

Copyright 2010 © Limsoon Wong

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

Copyright 2010 © Limsoon Wong

Any Question?

Acknowledgements

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

Copyright 2010 © Limsoon Wong

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

Copyright 2010 © Limsoon Wong

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

Copyright 2010 © Limsoon Wong

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

Copyright 2010 © Limsoon Wong