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
Unit 5: Sequence Homology Interpretation

Wong Limsoon
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
Brief recap of sequence comparison / alignment
Motivations for seq 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

<table>
<thead>
<tr>
<th>Amicyanin</th>
<th>MPHNVHFVAGVLGEAALKGPMKKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60  70  80  90  100</td>
</tr>
<tr>
<td>Ascorbate Oxidase</td>
<td>ILQRGTPWADGTASISQCAINPGETFYYNFTVDNPGTFFYHGHLGMQRSAGLYGSLI</td>
</tr>
<tr>
<td></td>
<td>70  80  90  100  110  120</td>
</tr>
</tbody>
</table>

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

```
>gil13476732|ref|NP_108301.1|
|gil14027493|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 MKPGLASIALAILFLPMAPPAHAATIOITMNELYISPTEVSASKVGDTIRWVNDVFAHT
       MK G L ++ MA PA AATIE+T++ LV SP V AKVGDTI WVN DV AHT
Subjct: 1 MKAGALIRLSWLAALALMAAAPAAATIEVTIDKLVFSPATVEAKVGDTIETWVNNDVVAHT

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| FHFTSWPDGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVVIDAMLD
gi|2499753| FHFTGWPDHGVPYHATGLLSFIIRVKLSNP--PSAGPIUVHCSAGAGRTGCYIVIDIMLD
gi|462550| YHTQWPDMGVPEYALPVTFTVRRSSAARM--PETGPVILHCSAGVGRTGTYIVIDSLQ
 gi|2499751| FHFTSWPDHGVPDTDLLINFRYLVDYMKVQSPPEPSPILHCSAGVGRGTGTFIAIDRLIY
gi|1709906| FQFTAUPDHVPEHPTPFLALRRVKTCPNP--PDAGPMVHVCSAGVGRTGCFIVIDAMLE
 gi|126471| LHFTSWPDGVPFTPIGMLKFLKKVKTLNP--VHAGPIUVHCSAGVGRGTGFVVIDAMMA
 gi|548626| FHFTGWPDHGVPYHATGLLSFIIRVKLSNP--PSAGPIUVHCSAGAGRTGCYIVIDIMLD
 gi|131570| FHFTGWPDHGVPYHATGLLGFRQVQVKSKSP--PNAGPLUVHCSAGAGRTGCFIVIDIMLD
 gi|2144715| FHFTSWPDHGVPDTDLLINFRYLVDYMKVQSPPEPSPILHCSAGVGRGTGTFIAIDRLIY
 ** *** *** . * . . . . . . **** ** . * .
```

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
- Proteins perform a wide variety of activities in the cell
Function assignment to protein seq

SPSTNRYYPPELPLEEEINRMLADDNKLFREEFNALPACPIQATCEAASKEENKEKNRYVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAAQGPKEETVNDFWRMIWEQNTATIVMTNLKERKECKCAQYWPDQGCTYGNVRVSVEDVTVLVDYTVRKFCIQQVDVTNRKPQRLITQFHFTSWPDGVPFTPIGMLKFLKKVAKCNPQYAGAVAHCSAGVGRTGTFVVIDAMLDMHESERKVDYGFVSRIRAQRQCMVQTDQYVFIFYQALHELHYLDTELET

• How do we attempt to assign a function to a new protein sequence?
In the course of evolution…
Remember this exercise?

Let \( a = AFPHQHRVP \)
Let \( b = PQVYNIMKE \)

Suppose each generation differs from the previous by 1 residue.

What is the max difference between the 2\(^{nd}\) generation of \( a \)?

What is the min difference between the 2\(^{nd}\) generation of \( a \) and \( b \)?
The triumph of logic

In the course of evolution...

Two proteins inheriting their function from a common ancestor have very similar amino acid sequences.
Exercise #1

How can we guess the function of a protein?
BLAST: How it works
Altschul et al., *JMB*, 215:403--410, 1990

- BLAST is one of the most popular tools for doing “guilt-by-association” sequence homology search

Exercise: Why do we need this step?
Homologs obtained by BLAST

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>14193729</td>
<td>gb</td>
<td>AAK56109.1</td>
</tr>
<tr>
<td>gi</td>
<td>126467</td>
<td>sp</td>
<td>P18433</td>
</tr>
<tr>
<td>gi</td>
<td>4506303</td>
<td>ref</td>
<td>NP_002827.1</td>
</tr>
<tr>
<td>gi</td>
<td>227294</td>
<td>prf</td>
<td>J1701300A</td>
</tr>
<tr>
<td>gi</td>
<td>18450369</td>
<td>ref</td>
<td>NP_543030.1</td>
</tr>
<tr>
<td>gi</td>
<td>32067</td>
<td>emb</td>
<td>CAA37447.1</td>
</tr>
<tr>
<td>gi</td>
<td>285113</td>
<td>pir</td>
<td>JC1285</td>
</tr>
<tr>
<td>gi</td>
<td>6981446</td>
<td>ref</td>
<td>NP_036895.1</td>
</tr>
<tr>
<td>gi</td>
<td>2098414</td>
<td>pdb</td>
<td>1YPO</td>
</tr>
<tr>
<td>gi</td>
<td>32313</td>
<td>emb</td>
<td>CAA38662.1</td>
</tr>
<tr>
<td>gi</td>
<td>450583</td>
<td>gb</td>
<td>AAB04150.1</td>
</tr>
<tr>
<td>gi</td>
<td>6679557</td>
<td>ref</td>
<td>NP_033006.1</td>
</tr>
<tr>
<td>gi</td>
<td>483922</td>
<td>gb</td>
<td>AAA17990.1</td>
</tr>
</tbody>
</table>

- Thus our example sequence could be a protein tyrosine phosphatase α (PTPα)
Example alignment with PTP_{\alpha}

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

Query: 1
SPSTNRKYPPPLPVDKLEEIEINRMADDNKLFFREEFNLPACPIQATCEAASXXXXXXXXR

Sbjct: 202
SPSTNRKYPPPLPVDKLEEIEINRMADDNKLFFREEFNLPACPIQATCEAASKEENKEKNR

Query: 61
YVINILPYDHSVHLTPVEGVPDSYINASFINGYQEKNKIAAQGPKHEETVNDFWRMIWE

Sbjct: 262
YVINILPYDHSVHLTPVEGVPDSYINASFINGYQEKNKIAAQGPKHEETVNDFWRMIWE

Query: 121
QNTATTIVMTNLKERKECKCAQYWPDCGWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD

Sbjct: 322
QNTATTIVMTNLKERKECKCAQYWPDCGWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD

Query: 181
VTNRKPQRLITQFHTSWPDFFGPFTPIGMLKFLKKVKACNPQYAGAIIVVHCSAGVGRTG

Sbjct: 382
VTNRKPQRLITQFHTSWPDFFGPFTPIGMLKFLKKVKACNPQYAGAIIVVHCSAGVGRTG

Query: 241
TFVVIDAMLDMohSERKVDVYGFVSRIRAQRQCMVQTMQYVFIVYQALLEHYLYGDTLE

Sbjct: 442
TFVVIDAMLDMohSERKVDVYGFVSRIRAQRQCMVQTMQYVFIVYQALLEHYLYGDTLE
Guilt by association: Caveats

• Ensure that the effect of database size has been accounted for

• Ensure that the function of the homolog 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
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: \( \frac{1}{365} = 0.3\% \)

• Q: What is the prob that there is a person in the room having the same birthday as you?
  • A: \( 1 - \left(\frac{364}{365}\right)^{365} = 63\% \)

• 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 $10^7 \times 10^{-6} = 10$ seqs in it that give an equally good alignment

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

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

Note: $P = 1 - e^{-E}$
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
• Alignment 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
Effect of sequence length

Examples of invalid function assignment: IMP dehydrogenases (IMPDH)

<table>
<thead>
<tr>
<th>ID</th>
<th>Organism</th>
<th>PIR</th>
<th>Swiss-Prot/TrEMBL</th>
<th>RefSeq/GenPept</th>
</tr>
</thead>
<tbody>
<tr>
<td>NF00181857</td>
<td>Methanococcus jannaschii</td>
<td>564381 conserved hypothetical protein MJ0653</td>
<td>Y653_METJA Hypothetical protein MJ0653</td>
<td>g1592300 inosine-5'-monophosphate dehydrogenase (guaB)</td>
</tr>
<tr>
<td>NF00187783</td>
<td>Archaeoglobus fulgidus</td>
<td>D69157 homolog AF0847 IMP dehydrogenase (guaB-1) homolog</td>
<td>C09411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)</td>
<td>g2649754 inosine monophosphate dehydrogenase (guaB)</td>
</tr>
<tr>
<td>NF00188267</td>
<td>Archaeoglobus fulgidus</td>
<td>D69114 yhcV homolog 2 IMP dehydrogenase (guaB-2) homolog</td>
<td>C08162 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-2)</td>
<td>g2642610 inosine monophosphate dehydrogenase (guaB)</td>
</tr>
<tr>
<td>NF00188697</td>
<td>Archaeoglobus fulgidus</td>
<td>D69157 homolog AF0847 IMP dehydrogenase (guaB-1) homolog</td>
<td>C09411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)</td>
<td>g2649754 inosine monophosphate dehydrogenase (guaB)</td>
</tr>
<tr>
<td>NF00197775</td>
<td>Thermococcus sp.</td>
<td>D69157 homolog AF0847 IMP dehydrogenase (guaB-1) homolog</td>
<td>C09411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)</td>
<td>g2649754 inosine monophosphate dehydrogenase (guaB)</td>
</tr>
<tr>
<td>NF00414709</td>
<td>Methanotroph fermentans</td>
<td>D69157 homolog AF0847 IMP dehydrogenase (guaB-1) homolog</td>
<td>C09411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)</td>
<td>g2649754 inosine monophosphate dehydrogenase (guaB)</td>
</tr>
<tr>
<td>NF00414811</td>
<td>Methanotroph fermentans</td>
<td>D69157 homolog AF0847 IMP dehydrogenase (guaB-1) homolog</td>
<td>C09411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)</td>
<td>g2649754 inosine monophosphate dehydrogenase (guaB)</td>
</tr>
<tr>
<td>NF00414837</td>
<td>Methanotroph fermentans</td>
<td>D69157 homolog AF0847 IMP dehydrogenase (guaB-1) homolog</td>
<td>C09411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)</td>
<td>g2649754 inosine monophosphate dehydrogenase (guaB)</td>
</tr>
<tr>
<td>NF00414969</td>
<td>Methanotroph fermentans</td>
<td>D69157 homolog AF0847 IMP dehydrogenase (guaB-1) homolog</td>
<td>C09411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)</td>
<td>g2649754 inosine monophosphate dehydrogenase (guaB)</td>
</tr>
</tbody>
</table>

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

Mis-assignment of function

No IMPDH domain
Emerging pattern

**Typical IMPDH**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Domain Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>A31907</td>
<td>PF000130: IMP dehydrogenase / GMP reductase signature</td>
</tr>
<tr>
<td>E70218</td>
<td>PF000130: IMP dehydrogenase / GMP reductase C terminus</td>
</tr>
<tr>
<td>E64381</td>
<td>PF01381: Helix-turn-helix</td>
</tr>
<tr>
<td>G69355</td>
<td>PF01381: Helix-turn-helix</td>
</tr>
<tr>
<td>F89514</td>
<td>PF01381: Helix-turn-helix</td>
</tr>
<tr>
<td>B69407</td>
<td>PF02195: FanB-like nuclease domain</td>
</tr>
</tbody>
</table>

**Functional IMPDH w/o CBS**

- IMPDH Misnomer in *Methanococcus jannaschii*
- IMPDH Misnomers in *Archaeoglobus fulgidus*

- 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
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 #2: Why?
In the course of evolution...
Multiple alignment of PTPs

- 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?
What if there is no useful seq homology?

- 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

- Genes (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

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

where

$$w_z = \binom{N}{z}$$
$$\overline{w}_z = \binom{N-z}{x-z} \cdot \binom{N-x}{y-z}$$
$$W = \binom{N}{x} \cdot \binom{N}{y}$$

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

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

- 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 #3: Why do proteins having high hamming distance also have this behaviour?
Guilt by association of dissimilarities

<table>
<thead>
<tr>
<th></th>
<th>Orange₁</th>
<th>Banana₁</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Apple₁</strong></td>
<td>Color = red vs orange</td>
<td>Color = red vs yellow</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Skin = smooth vs rough</td>
<td>Skin = smooth vs smooth</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Size = small vs small</td>
<td>Size = small vs small</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Shape = round vs round</td>
<td>Shape = round vs oblong</td>
<td>...</td>
</tr>
<tr>
<td><strong>Orange₂</strong></td>
<td>Color = orange vs orange</td>
<td>Color = orange vs yellow</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Skin = rough vs rough</td>
<td>Skin = rough vs smooth</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Size = small vs small</td>
<td>Size = small vs small</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Shape = round vs round</td>
<td>Shape = round vs oblong</td>
<td>...</td>
</tr>
<tr>
<td><strong>Unknown₁</strong></td>
<td>Color = red vs orange</td>
<td>Color = red vs yellow</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Skin = smooth vs rough</td>
<td>Skin = smooth vs smooth</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Size = small vs small</td>
<td>Size = small vs small</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Shape = round vs round</td>
<td>Shape = round vs oblong</td>
<td>...</td>
</tr>
</tbody>
</table>

“unknown” is an “apple”!
SVM-Pairwise framework

Training Data
S1
S2
S3
...

Feature Generation
Training Features
S1  S2  S3  ...
S1  f11  f12  f13  ...
S2  f21  f22  f23  ...
S3  f31  f32  f33  ...
...  ...  ...  ...  ...

Training
Support Vectors Machine
(Radial Basis Function Kernel)

Trained SVM Model
(Feature Weights)

Classification
RBF Kernel

Discriminant Scores

Testing Data
T1
T2
T3
...

Feature Generation
Testing Features
S1  S2  S3  ...
T1  f11  f12  f13  ...
T2  f21  f22  f23  ...
T3  f31  f32  f33  ...
...  ...  ...  ...  ...

Image credit: Kenny Chua

f_{31} is the local alignment score between S_3 and S_1

f_{31} is the local alignment score between T_3 and S_1
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

\[
PR = \frac{\sum_{i}^{K} k_i}{\sum_{i}^{K} m_i} \quad SN = \frac{\sum_{i}^{K} k_i}{\sum_{i}^{K} n_i}
\]

- \(n_i\) 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 \(i\)

\(\Rightarrow\) “level-2 only” neighbours performs better
\(\Rightarrow\) \(L1 \cap L2\) 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 btw two sets \( X \) and \( Y \)
- Greater weight given to similarity

⇒ **Similarity can be defined as**

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

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

<table>
<thead>
<tr>
<th>Neighbours</th>
<th>CD-Distance</th>
<th>FS-Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₁</td>
<td>0.471810</td>
<td>0.498745</td>
</tr>
<tr>
<td>S₂</td>
<td>0.224705</td>
<td>0.298843</td>
</tr>
<tr>
<td>S₁ ∪ S₂</td>
<td>0.224581</td>
<td>0.29629</td>
</tr>
</tbody>
</table>

- 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 between 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 between \(u\) and \(v\) is observed

<table>
<thead>
<tr>
<th>Source</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affinity Chromatography</td>
<td>0.823077</td>
</tr>
<tr>
<td>Affinity Precipitation</td>
<td>0.455904</td>
</tr>
<tr>
<td>Biochemical Assay</td>
<td>0.666667</td>
</tr>
<tr>
<td>Dosage Lethality</td>
<td>0.5</td>
</tr>
<tr>
<td>Purified Complex</td>
<td>0.891473</td>
</tr>
<tr>
<td>Reconstituted Complex</td>
<td>0.5</td>
</tr>
<tr>
<td>Synthetic Lethality</td>
<td>0.37386</td>
</tr>
<tr>
<td>Synthetic Rescue</td>
<td>1</td>
</tr>
<tr>
<td>Two Hybrid</td>
<td>0.265407</td>
</tr>
</tbody>
</table>
Exercise #4

Can you think of things a biologist can do to assess the overall reliability of a PPI screening assay / source?

<table>
<thead>
<tr>
<th>Source</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affinity Chromatography</td>
<td>0.823077</td>
</tr>
<tr>
<td>Affinity Precipitation</td>
<td>0.455904</td>
</tr>
<tr>
<td>Biochemical Assay</td>
<td>0.666667</td>
</tr>
<tr>
<td>Dosage Lethality</td>
<td>0.5</td>
</tr>
<tr>
<td>Purified Complex</td>
<td>0.891473</td>
</tr>
<tr>
<td>Reconstituted Complex</td>
<td>0.5</td>
</tr>
<tr>
<td>Synthetic Lethality</td>
<td>0.37386</td>
</tr>
<tr>
<td>Synthetic Rescue</td>
<td>1</td>
</tr>
<tr>
<td>Two Hybrid</td>
<td>0.265407</td>
</tr>
</tbody>
</table>
Functional similarity estimate: FS-weighted measure with reliability

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

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

• \( N_k \) 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 reliabilities

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

<table>
<thead>
<tr>
<th>Neighbours</th>
<th>CD-Distance</th>
<th>FS-Weight</th>
<th>FS-Weight R</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.471810</td>
<td>0.498745</td>
<td>0.532596</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.224705</td>
<td>0.298843</td>
<td>0.375317</td>
</tr>
<tr>
<td>$S_1 \cup S_2$</td>
<td>0.224581</td>
<td>0.29629</td>
<td>0.363025</td>
</tr>
</tbody>
</table>
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**

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

- \( r_{\text{int}} \) 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
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
  - Head of R&D at Data Robot
Application of sequence comparison: Key mutation site discovery
Identifying key mutation sites
K.L.Lim et al., JBC, 273:28986--28993, 1998

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?

Sequence from a typical PTP domain D2

>gi|00000|PTPA-D2
EEFKKLTSIKIQNDKMRGTPNLPAKKKRVRVLQIIIPYEFNIVIPVKRGEENTDYVNASFD
IDGYRQKDSYIASQGPLLHTIEDFWRMIVWFKSCSIIVMLTELEERGQKECAQYUHPDGLV
SYGDITVELKKEEEECSYTVRDLLVNTTRENKSRQIRQFHFHGWPEVGIISPDKGKMISII
AAVQKQQQQSGNHPITVHCSAGAGRTGTCALSTVLERVKAEGLDVFQTVKSLRLQRPH
MVQTLEQYECYKVQQEYIDAFSDYANFK
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

Exercise #5

Which of these two sites ("X" or "O") is more likely to explain the difference of D1 and D2?

- D2
  - Present: 
  - Absent: 

- D1
  - Present: 
  - Absent: 

absent

present
Key mutation site: PTP D1 vs D2

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

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

• 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?
About the inventor: Prasanna Kolatkar

- **Prasanna Kolatkar**
  - Research Fellow, BIC, NUS, 1997-1999
  - Currently Senior Scientist at Qatar Biomedical Research Institute
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


References

References

• H.N. Chua, W.-K. Sung. A better gap penalty for pairwise SVM. Proc. APBC05, pages 11-20

