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

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

Very Brief Recap of Sequence Comparison/Alignment

Sequence Alignment

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

Sequence Alignment: Poor Example

- Poor seq alignment shows few matched positions
  ⇒ The two proteins are not likely to be homologous

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

No obvious match between Amicyanin and Ascorbate Oxidase
Sequence Alignment: Good Example

- Good alignment usually has clusters of extensive matched positions
  \[ \Rightarrow \text{The two proteins are likely to be homologous} \]

Multiple Alignment: An Example

- Multiple seq alignment maximizes number of positions in agreement across several seqs
  \[ \Rightarrow \text{seqs belonging to same “family” usually have more conserved positions in a multiple seq alignment} \]

Function Assignment to Protein Sequence

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

- Compare the target sequence T with sequences S₁, ..., Sₙ of known function in a database
- Determine which ones amongst S₁, ..., Sₙ are the mostly likely homologs of T
- Then assign to T the same function as these homologs
- Finally, confirm with suitable wet experiments

BLAST: How It Works

- BLAST is one of the most popular tools for doing "guilt-by-association" sequence homology search
- Find seqs with good flanking alignment
- Find from db seqs with short perfect matches to query seq

Exercise: Why do we need this step?

Example Alignment with PTPα

Guilty-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
Law of Large Numbers

• Suppose you are in a room with 365 other people
• Q: What is the probability that a specific person in the room has the same birthday as you?
  • A: $1/365 = 0.3\%$

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

• Q: What is the probability 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 the probability 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: $P = 1 - e^{-k}$

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

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

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

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Organism</th>
<th>GO</th>
<th>Name</th>
<th>PubMed Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

IMPDH Misnomer in Methanococcus jannaschii

IMPDH Misnomer in Archaeoglobus fulgidus

• 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

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

Emerging Pattern

• 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

Application of Sequence Comparison: Active Site/Domain Discovery

• Easier if sequences of distance homologs are used

Exercise: 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 important with PTPs
- Else they wouldn’t be conserved by evolution
  ⇒ They are candidate active sites

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

- 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: P-value

The probability of observing by chance a co-occurrence of genes \( X \) and \( Y \) in a set of \( N \) lineages, given that \( X \) occurs in \( z \) lineages and \( Y \) in \( y \) lineages is

\[
P(\text{observed} | \text{null}) = \frac{\binom{z}{x} \binom{y}{y} \binom{N-z}{N-y}}{\binom{N}{N}}
\]

where

- \( \binom{n}{k} \) denotes the binomial coefficient, the number of ways to choose \( k \) elements from a set of \( n \) elements.

- \( z \) is the number of ways to distribute \( x \) co-occurrences over \( N \) lineages.

- \( y \) is the number of ways to distribute \( y \) co-occurrences over \( N \) lineages.

- \( N-z \) is the number of ways to distribute \( x+y \) co-occurrences over the remaining \( N-z \) lineages.

Phylogenetic Profiling: 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>Bacteriophage</td>
<td>69</td>
<td>197</td>
<td>27</td>
</tr>
<tr>
<td>Translation</td>
<td>30</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>RNA synthesis and processing</td>
<td>26</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Membrane transport</td>
<td>29</td>
<td>89</td>
<td>5</td>
</tr>
<tr>
<td>Fimbriae</td>
<td>20</td>
<td>89</td>
<td>3</td>
</tr>
<tr>
<td>Iron transport and metabolism</td>
<td>19</td>
<td>31</td>
<td>2</td>
</tr>
<tr>
<td>Glycine metabolism</td>
<td>18</td>
<td>31</td>
<td>2</td>
</tr>
<tr>
<td>Methionine and Methylthionine metabolism</td>
<td>12</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>1,064</td>
<td>148,226</td>
<td>8,408</td>
</tr>
</tbody>
</table>

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

Guilt by Association of Dissimilarities

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

SVM-Pairwise Framework

Training Data

\[ S_1, S_2, S_3, \ldots \]

Testing Data

\[ T_1, T_2, T_3, \ldots \]

Feature Generation

Training Features

\[ f_{11}, f_{12}, f_{13}, \ldots, f_{1n} \]

Testing Features

\[ f_{21}, f_{22}, f_{23}, \ldots, f_{2n} \]

Feature Weights

Support Vectors Machine (Radial Basis Function Kernel)

Discriminant Scores

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?
- **SH3 Proteins**
- **SH3-Binding Proteins**
  - YW032c
  - Yoc667
  - Ydc84
  - Ypr138w
  - Ypl514w
  - Yal994w

  - 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

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

Functional Similarity Estimate: Czekanowski-Dice Distance
- **Functional distance between two proteins** (Brun et al, 2003)
  \[
  D(u, v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v|}
  \]
  - \(N_u\) 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

  \[ S(u, v) = 1 - D(u, v) = \frac{2X}{2X + (Y + Z)} \]
Functional Similarity Estimate: FS-Weighted Measure

- **FS-weighted measure**
  \[ S(u,v) = \frac{4N_v \cap N_u}{|N_v| - |N_u| + 2|N_v \cap N_u|} \times \frac{4N_u \cap N_v}{|N_u| - |N_v| + 2|N_u \cap N_v|} \]

  - \( N_v \) is the set of interacting partners of \( k \)
  - Greater weight given to similarity

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

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

Correlation w/ Functional Similarity

- **Correlation betw functional similarity & estimates**

<table>
<thead>
<tr>
<th>Neighbours</th>
<th>L1-Weight</th>
<th>L2-Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_1 )</td>
<td>0.497345</td>
<td>0.497345</td>
</tr>
<tr>
<td>( S_2 )</td>
<td>0.298343</td>
<td>0.298343</td>
</tr>
<tr>
<td>( S_1 \cup S_2 )</td>
<td>0.294511</td>
<td>0.294511</td>
</tr>
</tbody>
</table>

Reliability of Expt Sources

- **Diff Expt Sources have diff reliabilities**
  - Assign reliability to an interaction based on its expt sources
  - **Reliability betw u and v**

\[ r_{u,v} = 1 - \prod_{i \in R_{u,v}} (1-r_i) \]

- \( r_i \) is reliability of expt source \( i \)
- \( R_{u,v} \) is the set of expt sources in which interaction betw u and v is observed

Integrating Reliability

- **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>0.497345</td>
<td>0.532906</td>
</tr>
<tr>
<td>( S_2 )</td>
<td>0.298343</td>
<td>0.298343</td>
<td>0.298343</td>
</tr>
<tr>
<td>( S_1 \cup S_2 )</td>
<td>0.294511</td>
<td>0.294511</td>
<td>0.294511</td>
</tr>
</tbody>
</table>

Functional Similarity Estimate: FS-Weighted Measure with Reliability

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

\[ S(u,v) = \frac{2 \sum_{i \in R_{u,v}} r_i}{\left( \sum_{i \in R_{u,v}} r_i \right)^{1/2} \left( \sum_{j \in R_{v,u}} r_j \right)^{1/2}} \times \frac{2 \sum_{i \in R_{v,u}} r_i}{\left( \sum_{i \in R_{v,u}} r_i \right)^{1/2} \left( \sum_{j \in R_{u,v}} r_j \right)^{1/2}} \]

- \( N_v \) is the set of interacting partners of \( k \)
- \( r_{u,v} \) is reliability weight of interaction betw u and v

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

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**

\[
\hat{f}(u) = \frac{1}{Z} \left[ \alpha_r \sigma_r + \sum_{x \in \mathcal{N}_u} S_{r,x}(u,v) \delta(v,x) + \sum_{x \in \mathcal{N}_u} S_{r,x}(u,w) \delta(w,x) \right]
\]

- \( \alpha_r \) is fraction of all interaction pairs sharing function
- \( \sigma_r \) is weight of contribution of background freq
- \( \delta(k,x) = 1 \) if \( k \) has function \( x \), 0 otherwise
- \( \mathcal{N}_u \) is the set of interacting partners of \( u \)
- \( \sigma_r \) is freq of function \( x \) in the dataset
- \( Z \) is sum of all weights

```
Z = 1 + \sum_{x \in \mathcal{N}_u} S_{r,x}(u,v) + \sum_{x \in \mathcal{N}_u} S_{r,x}(u,w)
```

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

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 is 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

Key Mutation Site: PTP D1 vs D2

<table>
<thead>
<tr>
<th>Position</th>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>!</td>
<td>?</td>
</tr>
</tbody>
</table>

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

Confirnation by Mutagenesis Exp

• 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 exp?

About the Inventor: Prasanna Kolatkar

• Prasanna Kolatkar
  – Research Fellow, BIC, NUS, 1997-1999
  – Currently Group Leader at GIS
Concluding Remarks

What have we learned?

• General methodologies & applications
  – Guilt by association for protein function inference
  – Invariants for active site discovery
  – Emerging patterns for mutation site discovery

• Important tactics
  – Genome phylogenetic profiling
  – SVM-Pairwise
  – Protein-protein interactions

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
• L.J.Jensen et al. "Prediction of human protein function from post-translational modifications and localization features", JMB, 319:1257–1265, 2002
References