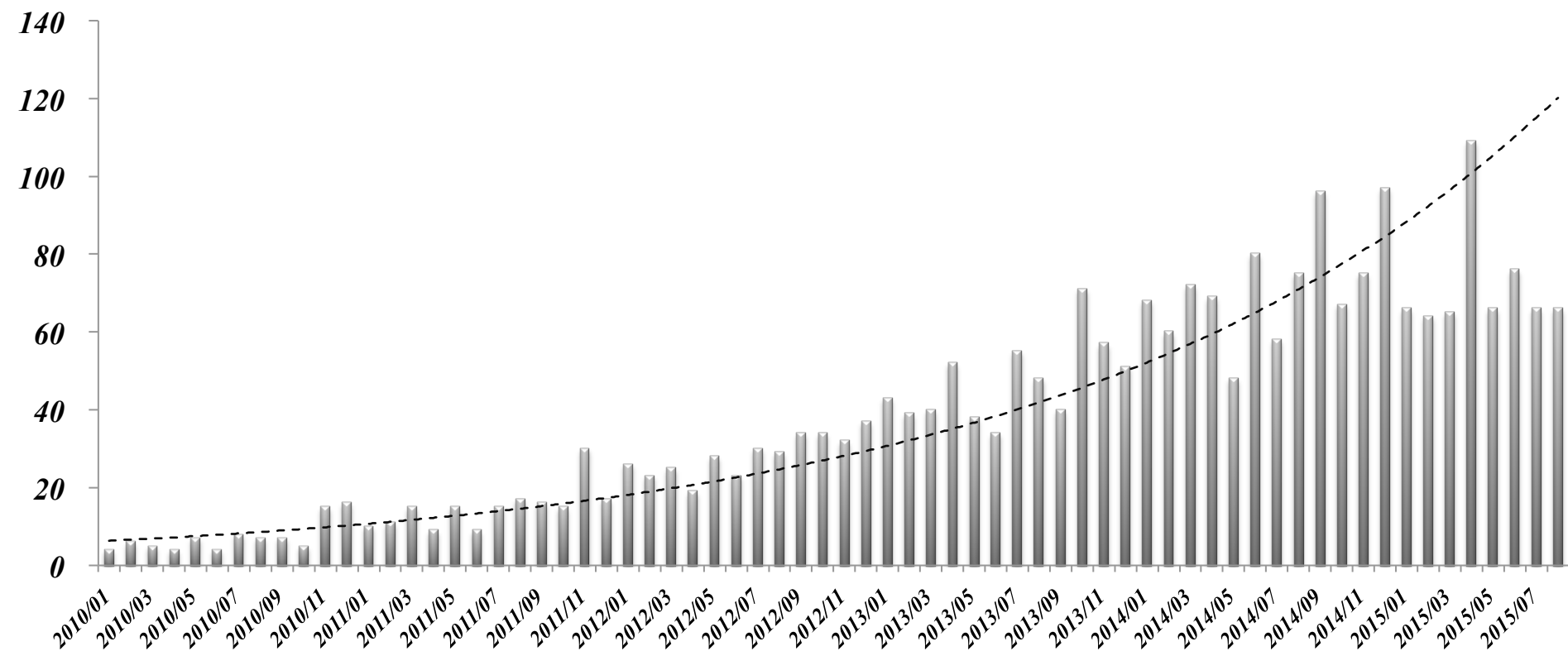


Why Orthofuzz ?



The search results of the term “RNA Seq transcriptome”: The results grouped based on date

Three issues with examples

- Same contig mapping relatively weakly to several families.
- Multiple contigs from the same assembly that are very close to each other sharing reads mapping and affecting expression levels.
- Many – Many mapping makes it impossible to ask questions experimenters are interested in?
 - Is an enzyme x expressed in the data?
 - Is a gene expressed in the system?

Illustration

- 1. Fragmentation
- 2. Many – Many mapping
- 3. Mismapped reads causing expression problem.

- Figures here

FRAGMENTATION

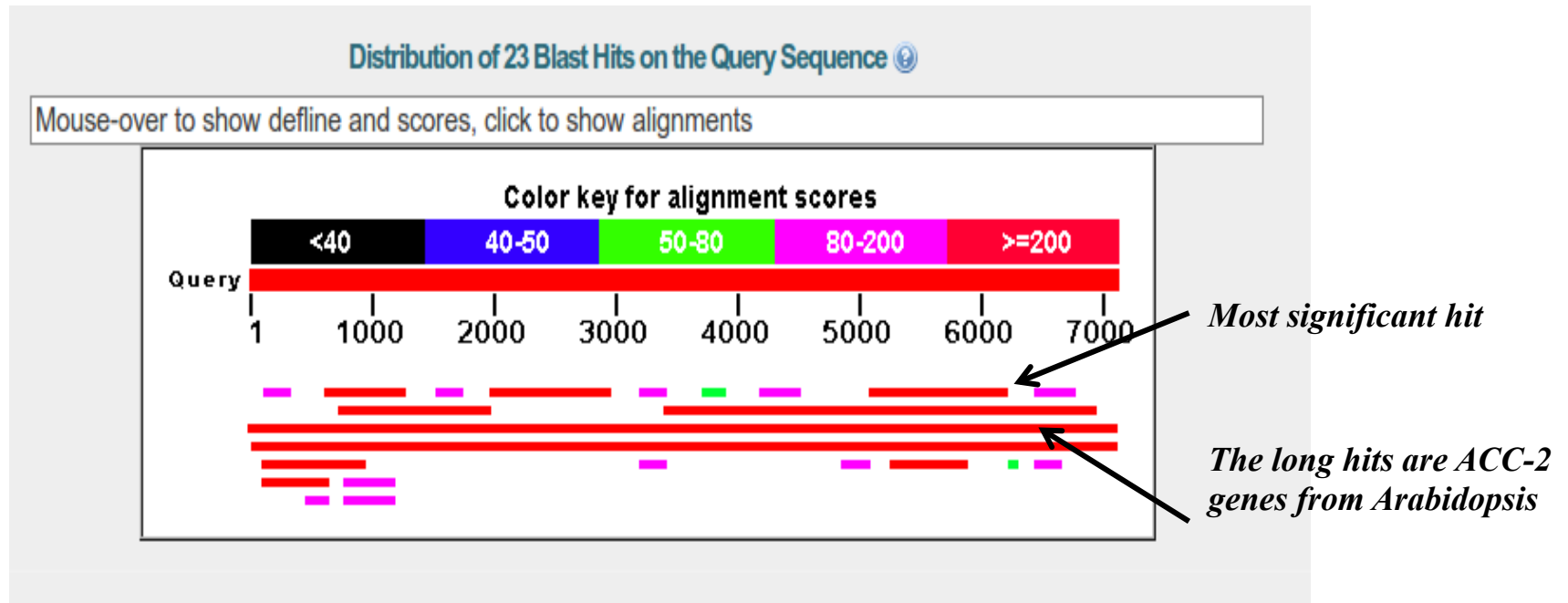
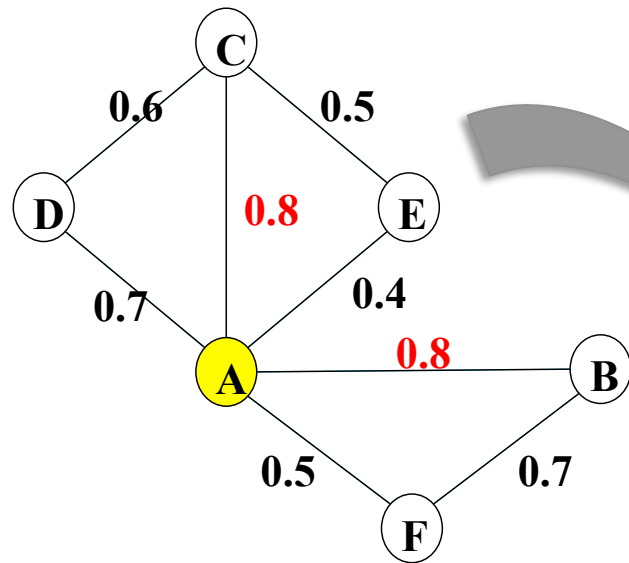


Figure 1: Example of fragmented contigs - The homologs of Acetyl CoA Carboxylase (AT1G36160.1) from *Arabidopsis thaliana* visualized on the NCBI BLAST server. The contigs are obtained from transcriptome assembled *de-novo* from Corn (B-73) Silks. Contigs of varying bit score and coverage show significant hits. The Contigs were assembled using Trinity Transcriptome Assembler²

Network formulation

Random Walking with Restart !

$$p^{t+1} = (1 - r)W_{p^t} + rp^0$$



$$p_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad W_{p^0} =$$

	A	B	C	D	E	F
A	0	0.8	0.8	0.7	0	0.5
B	0.8	0	0	0	0	0.7
C	0.8	0	0	0.6	0.5	0
D	0.7	0	0.6	0	0	0
E	0.4	0	0.5	0	0	0
F	0.5	0.7	0	0	0	0

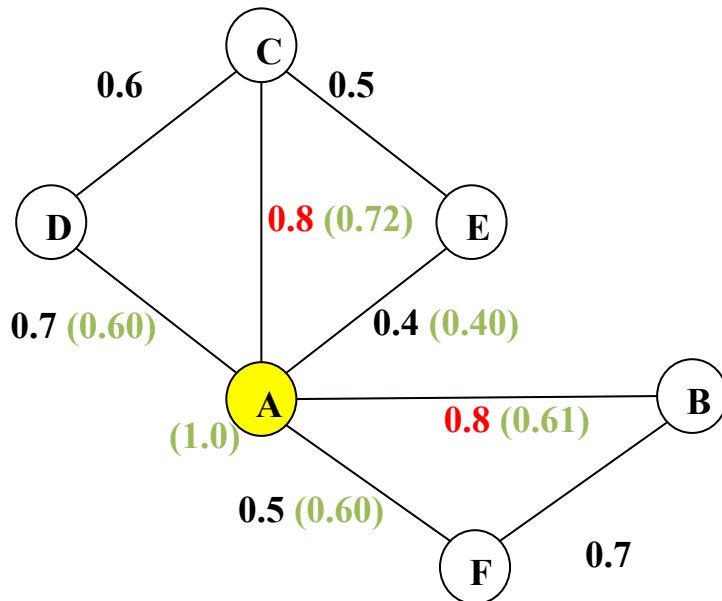
Run RWR:

tolerance = 0.06

Max Iterations = 1000

$r = 0.33$

Normalize by maximum score



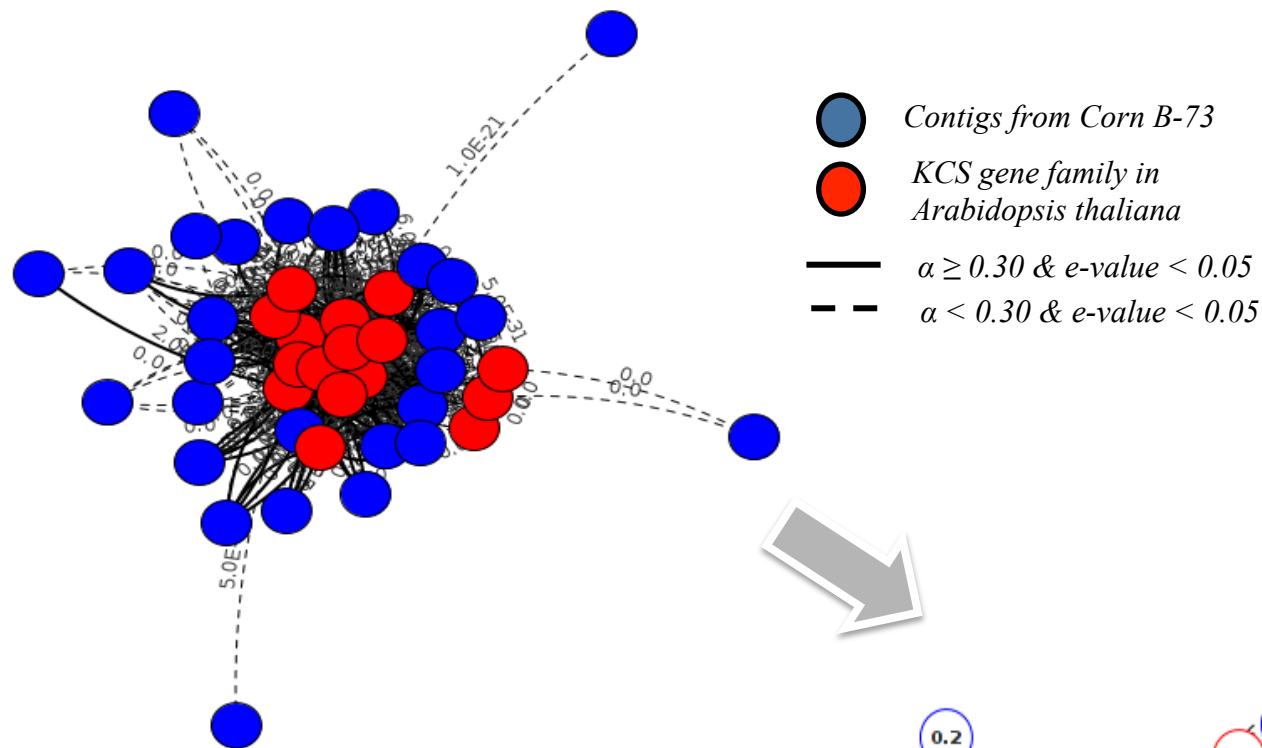


Figure : Only edges with $e\text{-value} < 0.05$ are shown in the graph. Even though, all sequences have significant hits, only some share multiple neighbors (multiple dark edges to the query family) with each other.

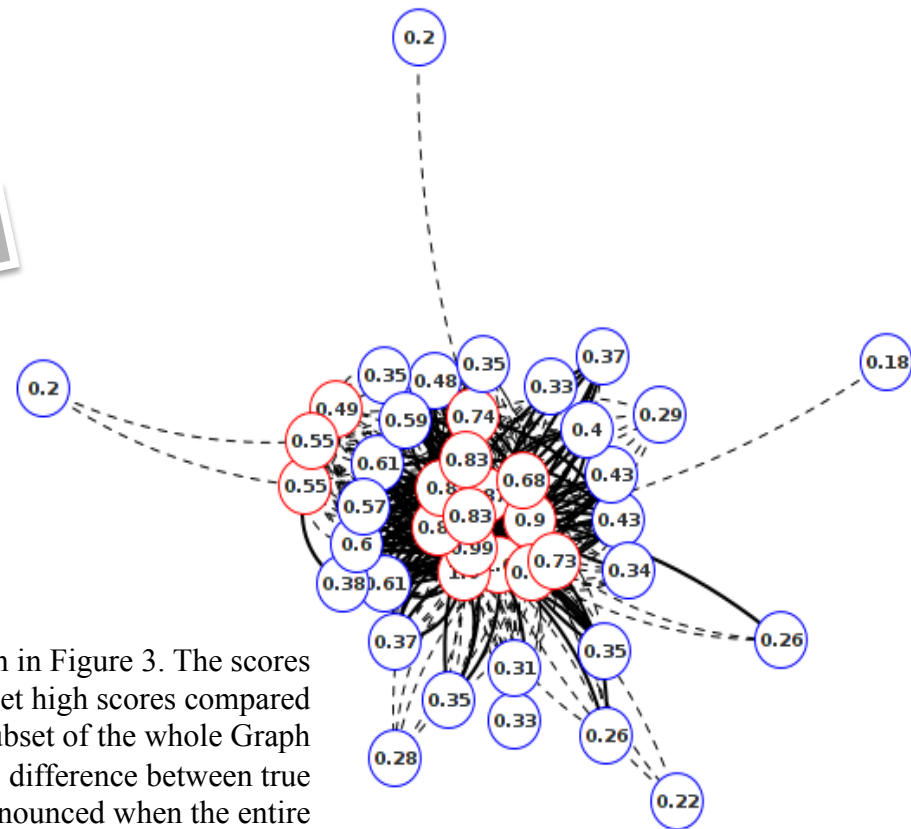


Figure : Result of the algorithm on the Graph in Figure 3. The scores are shown in the nodes. The true homologs get high scores compared to weaker hits. (Note that this network is a subset of the whole Graph and is for illustration purpose only. The score difference between true homologs and others becomes more pronounced when the entire graph is used.)

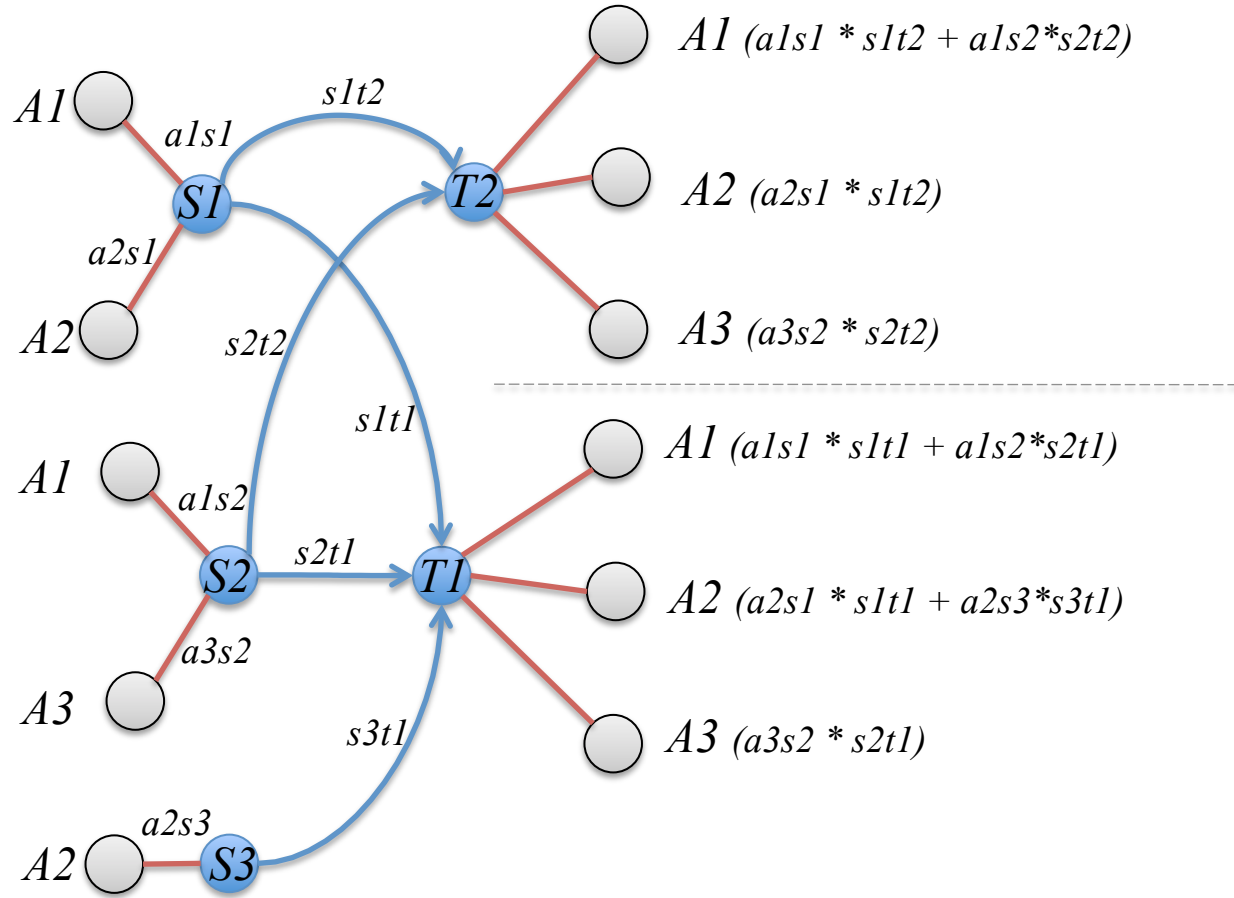
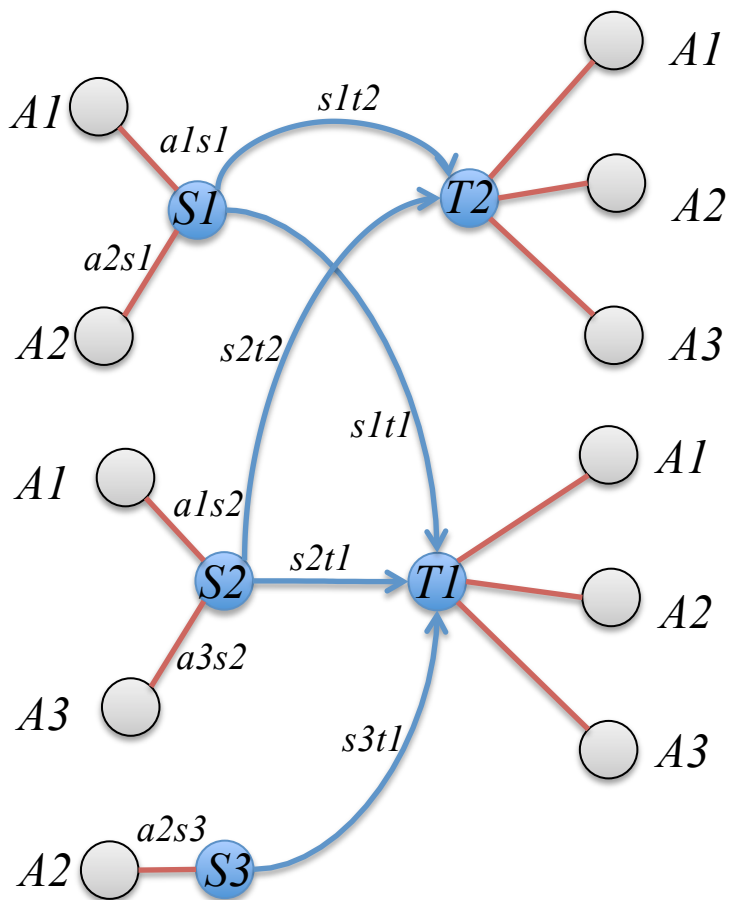


Figure X: Annotation flow network: A1-A3 are annotations of source sequences S1, S2 and S3. $axsy$ represents the user-defined confidence of the annotation ax to be associated with sequence sy . $sxty$ is the maximum score normalized orthofuzz score(sx, ty) obtained by querying the pairwise sequence similarity network using sx .



*This translates as
matrix multiplication of
Annotation Confidence Matrix
and Orthofuzz matrix*

$$\begin{array}{c}
 \begin{matrix} SS1 & SS2 & SS3 \\ A1 & \begin{pmatrix} a1s1 & a1s2 & 0 \end{pmatrix} \\ A2 & \begin{pmatrix} a2s1 & 0 & a2s3 \end{pmatrix} \\ A3 & \begin{pmatrix} 0 & a3s2 & 0 \end{pmatrix} \end{matrix} \times \begin{matrix} TS1 & TS2 \\ SS1 & \begin{pmatrix} s1t1 & s1t2 \end{pmatrix} \\ SS2 & \begin{pmatrix} s2t1 & s2t2 \end{pmatrix} \\ SS3 & \begin{pmatrix} s3t1 & 0 \end{pmatrix} \end{matrix} = \begin{matrix} TS1 & TS2 \\ A1 & \begin{pmatrix} a1s1 \times s1t1 + a1s2 \times s2t1 & a1s1 \times s1t2 + a1s2 \times s2t2 \end{pmatrix} \\ A2 & \begin{pmatrix} a2s1 \times s1t1 + a2s3 \times s3t1 & a2s1 \times s1t2 \end{pmatrix} \\ A3 & \begin{pmatrix} a3s2 \times s2t1 & a3s2 \times s2t2 \end{pmatrix} \end{matrix}
 \end{array}$$

$$\begin{array}{c}
\begin{array}{ccc}
SS1 & SS2 & SS3 \\
A1 & \begin{pmatrix} a1s1 & a1s2 & 0 \end{pmatrix} \\
A2 & \begin{pmatrix} a2s1 & 0 & a2s3 \end{pmatrix} \\
A3 & \begin{pmatrix} 0 & a3s2 & 0 \end{pmatrix}
\end{array}
\times
\begin{array}{cc}
\begin{array}{cc}
TS1 & TS2 \\
SS1 & \begin{pmatrix} s1t1 & s1t2 \end{pmatrix} \\
SS2 & \begin{pmatrix} s2t1 & s2t2 \end{pmatrix} \\
SS3 & \begin{pmatrix} s3t1 & 0 \end{pmatrix}
\end{array}
\end{array}
=
\begin{array}{cc}
\begin{array}{cc}
TS1 & TS2 \\
A1 & \begin{pmatrix} a1s1 \times s1t1 + a1s2 \times s2t1 & a1s1 \times s1t2 + a1s2 \times s2t2 \end{pmatrix} \\
A2 & \begin{pmatrix} a2s1 \times s1t1 + a2s3 \times s3t1 & a2s1 \times s1t2 \end{pmatrix} \\
A3 & \begin{pmatrix} a3s2 \times s2t1 & a3s2 \times s2t2 \end{pmatrix}
\end{array}
\end{array}
\end{array}$$

$$\mathbf{ATS}_{N_A \times N_{TS}} = \mathbf{ASS}_{N_A \times N_{SS}} \times \mathbf{SSTS}_{N_{SS} \times N_{TS}}$$

Symbol	Description
SS	<i>Set of Source Sequences</i>
N_{SS}	<i>Total Number of Source Sequences</i>
A	<i>Set of Source Annotations</i>
N_A	<i>Total Number of Annotations from Source Sequences</i>
TS	<i>Set of Target Sequences</i>
N_{TS}	<i>Total Number of Target Sequences</i>
ATS	<i>Target Annotation Weight Matrix</i>
ASS	<i>Source Annotation Confidence Matrix</i>
$SSTS$	<i>Source Target Orthofuzz Matrix</i>
$axsy$	<i>Confidence of assigning ax to sy</i>
$sytz$	<i>Max. normalized orthofuzz score of sy to tz</i>
$axtz$	<i>Annotation weight of ax to tz</i>

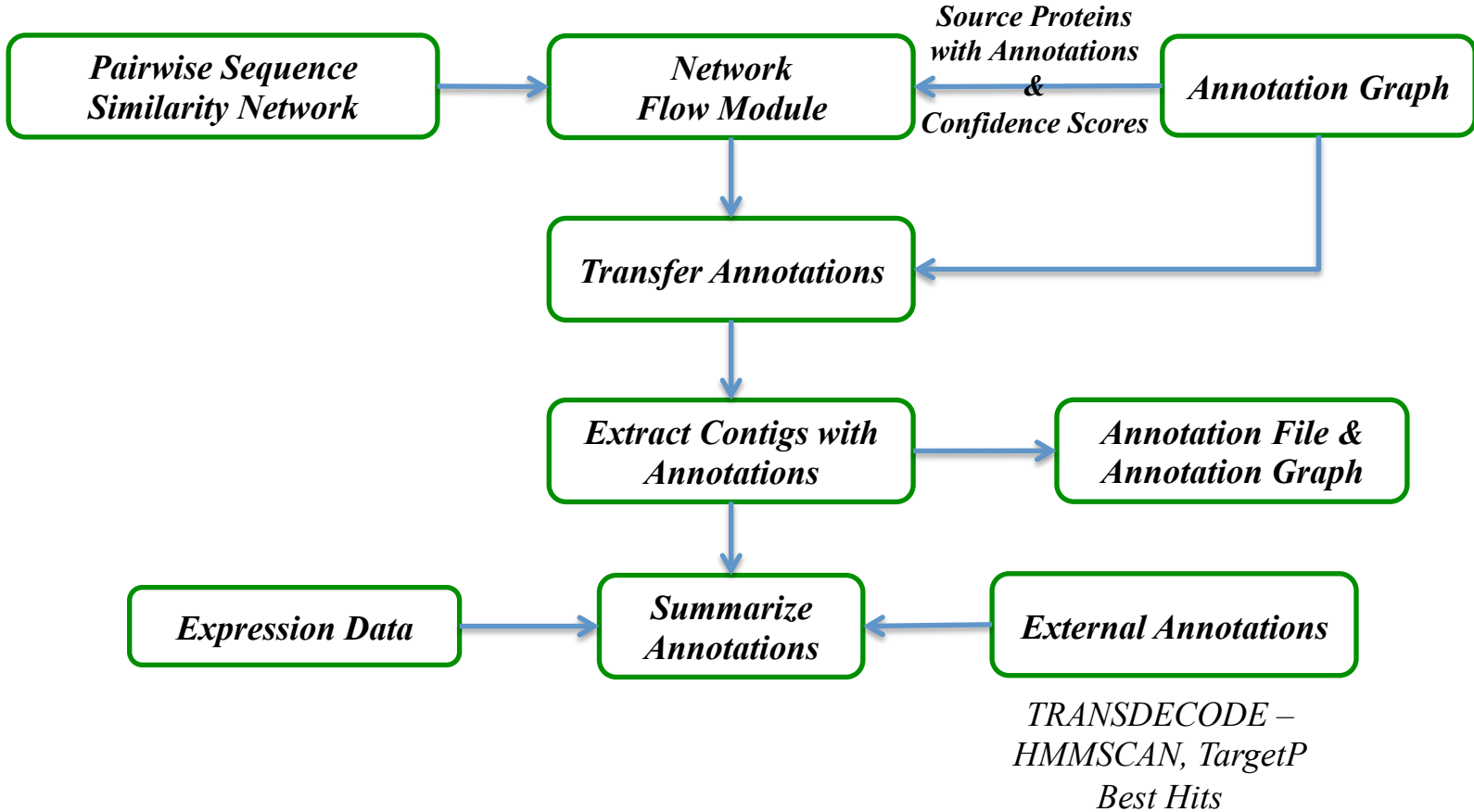


Figure X: Protocol used for assigning functional annotations to the de-novo assembled contigs

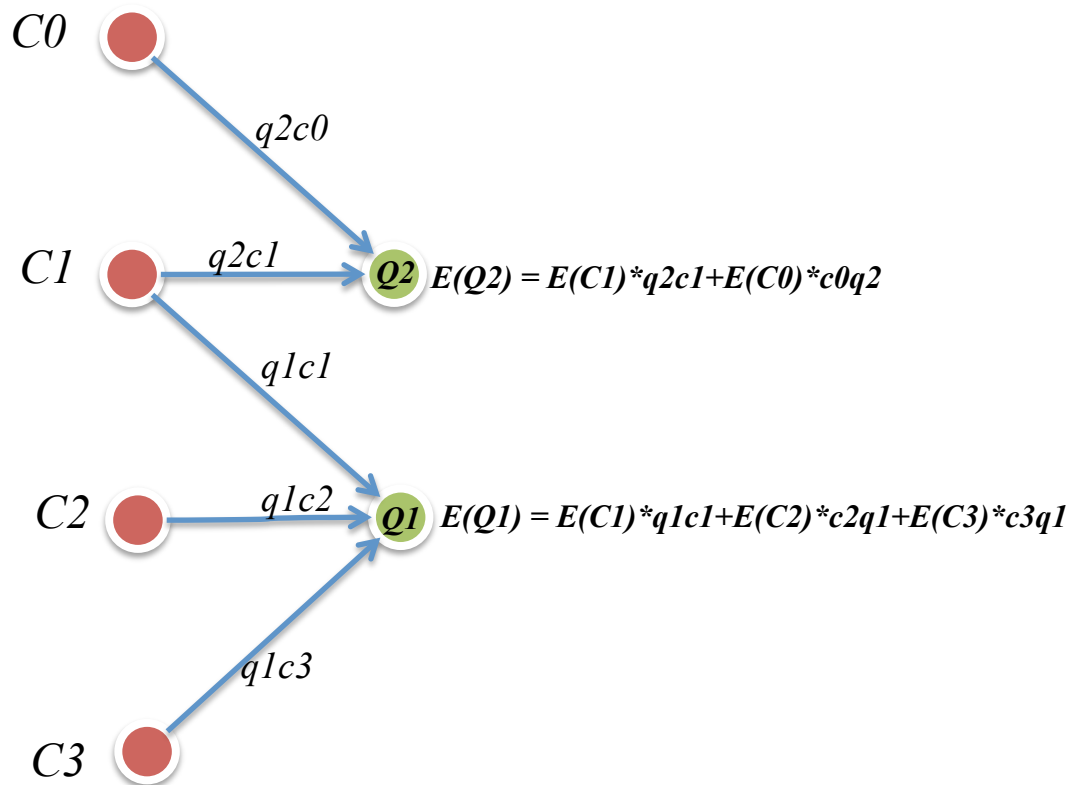
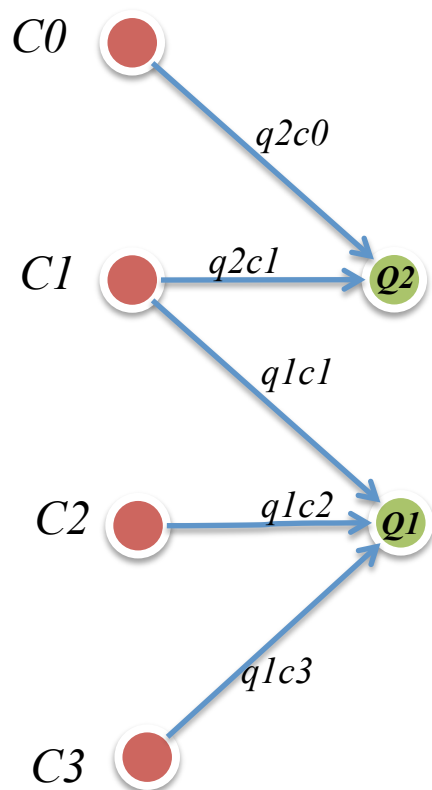


Figure X: Expression summary network: $C0$ - $C3$ are the contigs identified as homologs of Query Sets $Q1$ and $Q2$. $E(X)$ is the normalized expression level of X . q_{xyc} is the within species normalized orthofuzzscore ($q_{x,c}$) obtained by querying the network using the query set Q .



*This translates as
matrix multiplication of
OrthoFuzz Matrix and
Contig Expression Vector*

$$\begin{matrix} & C0 & C1 & C2 & C3 \\ Q1 & \begin{pmatrix} 0 & q1c1 & q1c2 & q1c3 \end{pmatrix} \\ Q2 & \begin{pmatrix} q2c0 & q2c1 & 0 & 0 \end{pmatrix} \end{matrix} \times \begin{matrix} E \\ C0 \begin{pmatrix} e_{c0} \\ e_{c1} \\ e_{c2} \\ e_{c3} \end{pmatrix} \end{matrix} = \begin{matrix} E \\ Q1 \begin{pmatrix} e_{c1} \times q1c1 + e_{c2} \times q1c2 + e_{c3} \times q1c3 \\ e_{c0} \times q2c0 + e_{c1} \times q2c1 \end{pmatrix} \end{matrix}$$

$$\begin{matrix} & C0 & C1 & C2 & C3 \\ Q1 \left(\begin{matrix} 0 & q1c1 & q1c2 & q1c3 \end{matrix} \right) \\ Q2 \left(\begin{matrix} q2c0 & q2c1 & 0 & 0 \end{matrix} \right) \end{matrix} \times \begin{matrix} E \\ C0 \left(\begin{matrix} e_{c0} \\ e_{c1} \\ e_{c2} \\ e_{c3} \end{matrix} \right) \end{matrix} = \begin{matrix} E \\ Q1 \left(\begin{matrix} e_{c1} \times q1c1 + e_{c2} \times q1c2 + e_{c3} \times q1c3 \\ e_{c0} \times q2c0 + e_{c1} \times q2c1 \end{matrix} \right) \end{matrix}$$

$$\mathbf{QC}_{N_Q \times N_C} \times \mathbf{Ce}_{N_C \times 1} = \mathbf{Qe}_{N_Q \times 1}$$

Symbol	Description
Q	<i>Set of Query Sets</i>
N_Q	<i>Total Number of Query Sets</i>
N_C	<i>Total Number of Expressed Contigs</i>
QC	<i>Query Contig Orthofuzz Matrix</i>
$qxcy$	<i>Species.normalized orthofuzz score of qx to cy</i>
Ce	<i>Vector containing expression values of Contigs</i>
e_{cy}	<i>Expression value of contig y</i>
Qe	<i>Vector containing expression values of Query</i>

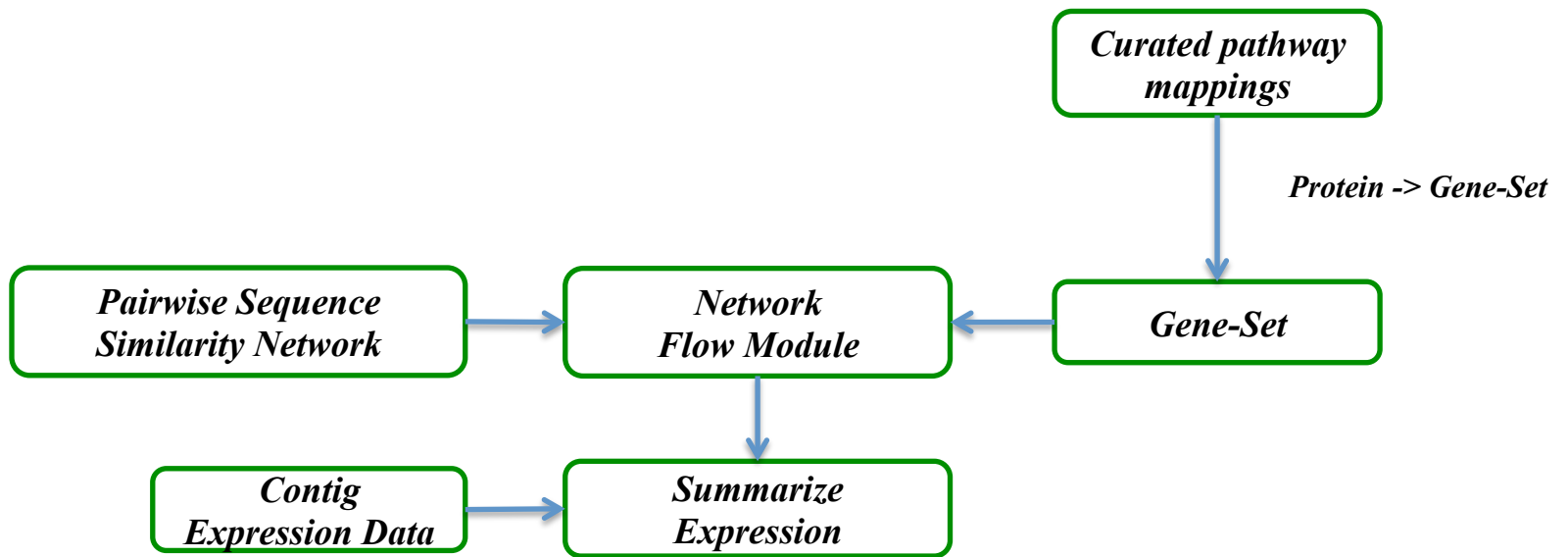
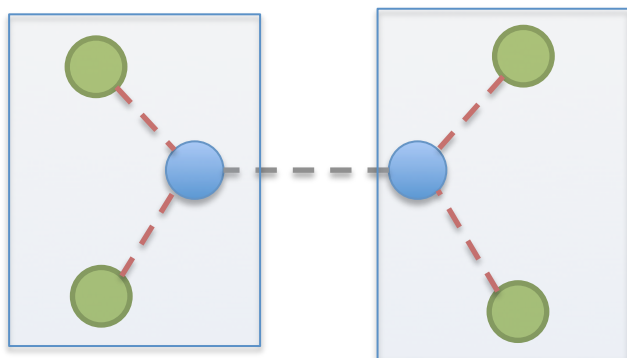
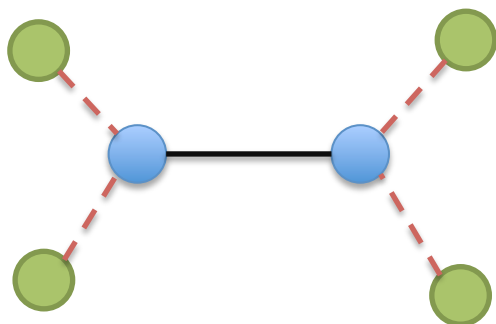
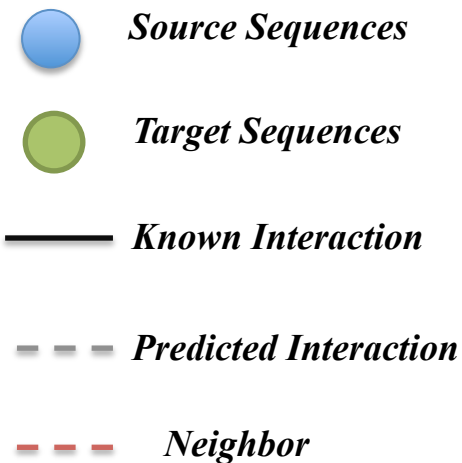


Figure: Protocol used for estimating expression levels of gene-sets



Is there evidence for protein complexes ?



Expression Levels

Interaction Unit 1

Interaction Unit 2

<i>Conditions</i>		



Are there interaction units whose expression levels are changing ?

Are they co-expressed ?

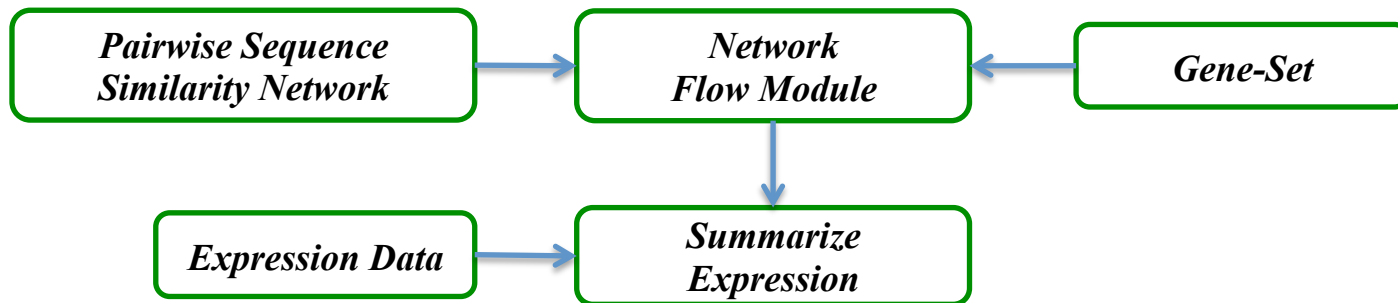


Figure: Protocol used for estimating expression levels of gene-sets

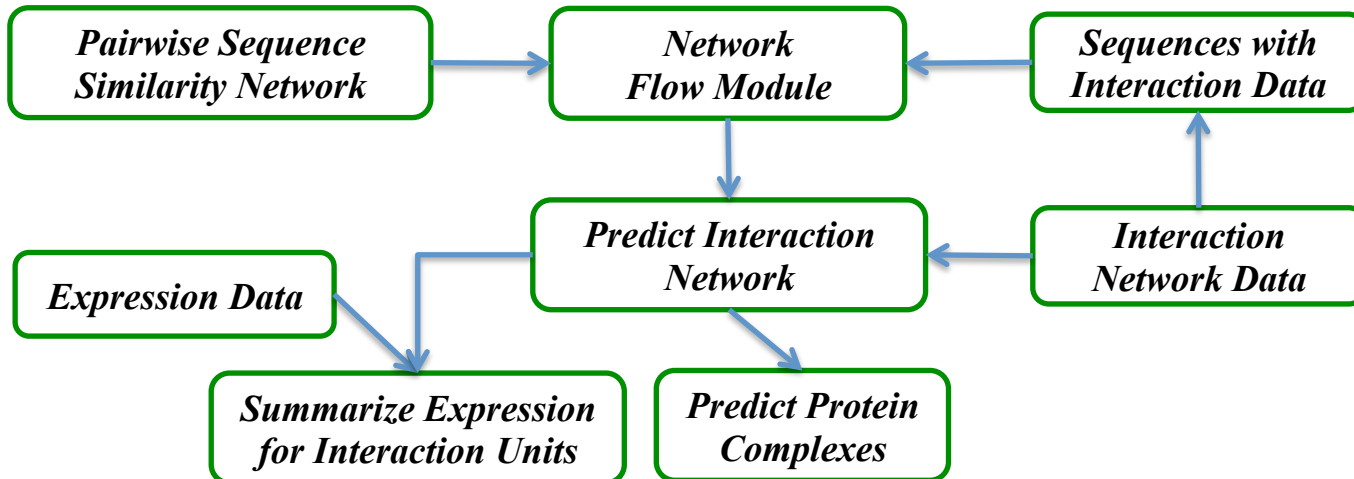
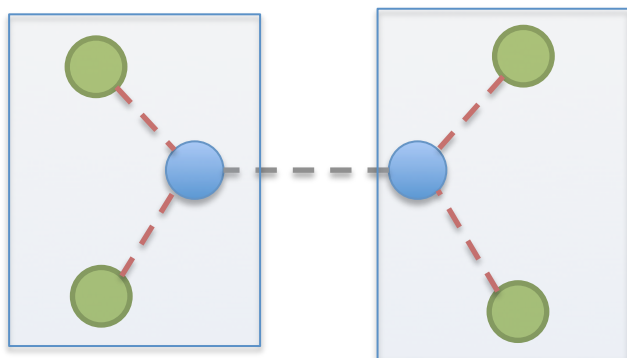
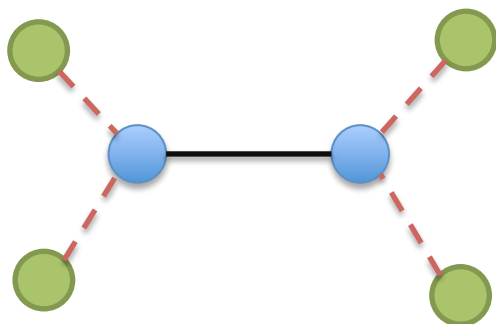
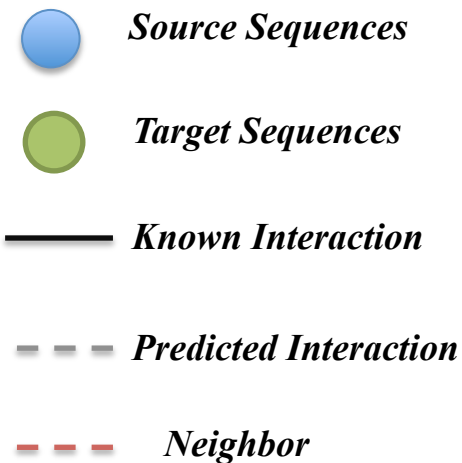


Figure: Protocol used for estimating expression levels of gene-sets



Is there evidence for protein complexes ?



Expression Levels

Interaction Unit 1

Interaction Unit 2

<i>Conditions</i>		



Are there interaction units whose expression levels are changing ?

Are they co-expressed ?