```
                                                    fastx toolkit      ┌─────────────────┐
                            ┌──────────────────────┐ ←──────────────── │  Illumina PE    │
                            │ Filtered Short Reads  │                   │ Transcriptome   │
                            └──────────────────────┘                   │  raw data       │
                                                                        └─────────────────┘
                                    Map using Bowtie

Pool All unmapped reads      ┌──────────────────────────┐
                             │ Reference Transcript Models[1] │
Trinity          Map using Bowtie
Assembler                            Estimate FPKM/TMM
                                     values using RSEM
   ┌──────────────────────────┐  Estimate FPKM/TMM
   │ Denovo assembled Contigs │  values using RSEM
   └──────────────────────────┘

Transdecode,
HMMSCAN &                      ┌──────────────────────────┐
TargetP                        │ Normalized Expression values │
                               └──────────────────────────┘

                               ┌──────────────────────────┐   ┌─────────────────────┐
                               │   ORTHOFUZZ package       │ ← │ SOURCE ANNOTATIONS  │
                               └──────────────────────────┘   │ PATHWAY MAPPINGS    │
                                                              │ SOURCE INTERACTIONS │
                                                              └─────────────────────┘
```

ORTHOFUZZ package

┌──────────────────────────────────┐   ┌──────────────────────────────────────┐
│ Functional Annotations: from   │   │ Interaction Data: │
│ ORTHOFUZZ & Transdecode   │   │ Predicted Co-expressed Interaction   │
└──────────────────────────────────┘   │ Complexes with Enrichment Statistics │
                                               └──────────────────────────────────────┘

┌──────────────────────────────────┐
│ Pathway Data: │
│ Pathway Enrichment Statistics   │
│ Expression of Enzymes in Pathways │
└──────────────────────────────────┘

┌──────────────────────────────────┐   ┌──────────────────────────────────────┐
│ List Enrichment Results: │   │ Hierarchical Clustering of Enzyme   │
│ Functional terms enriched in selected │   │ Expressions in a Heat Map   │
│ lists of transcripts/contigs   │   └──────────────────────────────────────┘
└──────────────────────────────────┘

Select source organims

Organism Name	Number of Contigs
Coccomyxasubellipsoidea C-169	16644
Chlorella variabilis	13342
Volvox carteri	12252
Chlamydomonas reinhardtii	11459
Glycine max	11450
Populus trichocarpa	11289
Setaria italica	11257
Physcomitrella patens subsp. patens	11253
Vitis vinifera	11226
Brachypodium distachyon	11208
Solanum lycopersicum	11056
Arabidopsis thaliana	11045
Brassica rapa subsp. pekinensis	11030
Selaginella moellendorffii	11017
Cicer arietinum	10982
Fragaria vesca subsp. vesca	10966
Cucumis sativus	10964
Prunus persica	10909
Ricinus communis	10800

Random Walking with Restart !

$$p^{t+1} = (1-r)W_{p^t} + rp^0$$
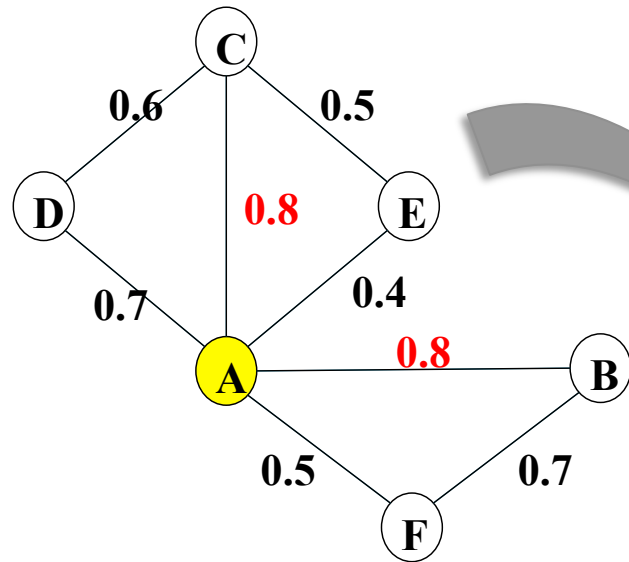
$$p_0 = \begin{vmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{vmatrix} \quad W_{p^0} =$$

	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:
tolerence = 0.06
Max Iterations = 1000

$$r = 0.33$$

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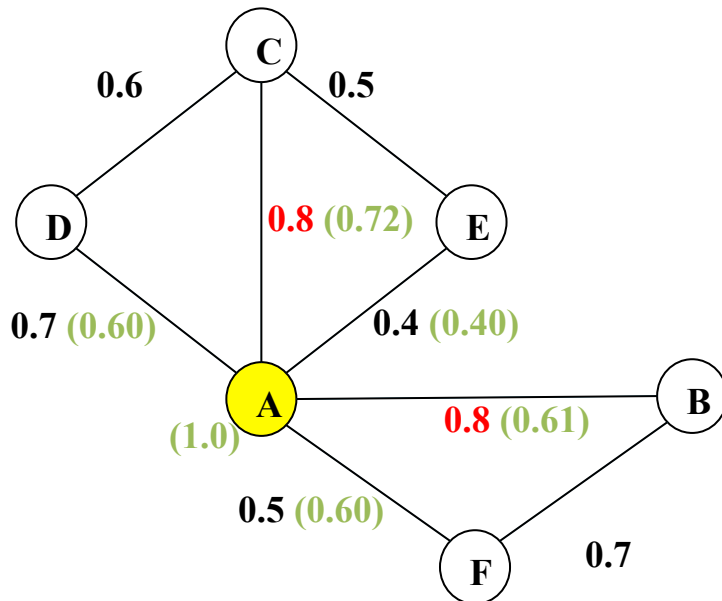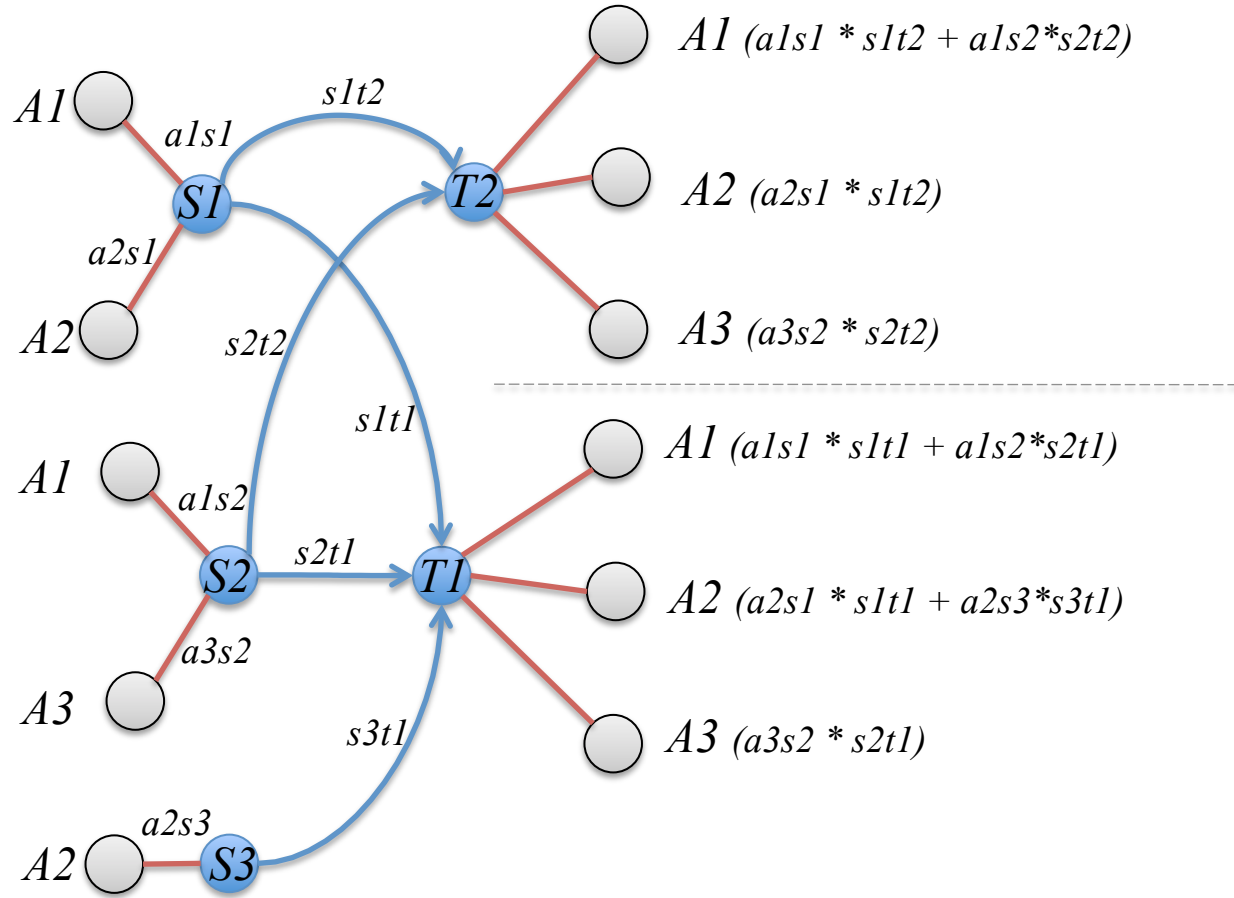
Figure : Annotation flow network: A1-A3 are annotations of source sequences S1, S2 and S3. axsy represents the user-defined confidence of the annotation ax to be associated with sequence sy. sxty is the maximum score normalized orthofuzz score(sx,ty) obtained by querying the pairwise sequence similarity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

A1 (a1s1 * s1t2 + a1s2*s2t2)

A2 (a2s1 * s1t2)

A3 (a3s2 * s2t2)

A1 (a1s1 * s1t1 + a1s2*s2t1)

A2 (a2s1 * s1t1 + a2s3*s3t1)

A3 (a3s2 * s2t1)
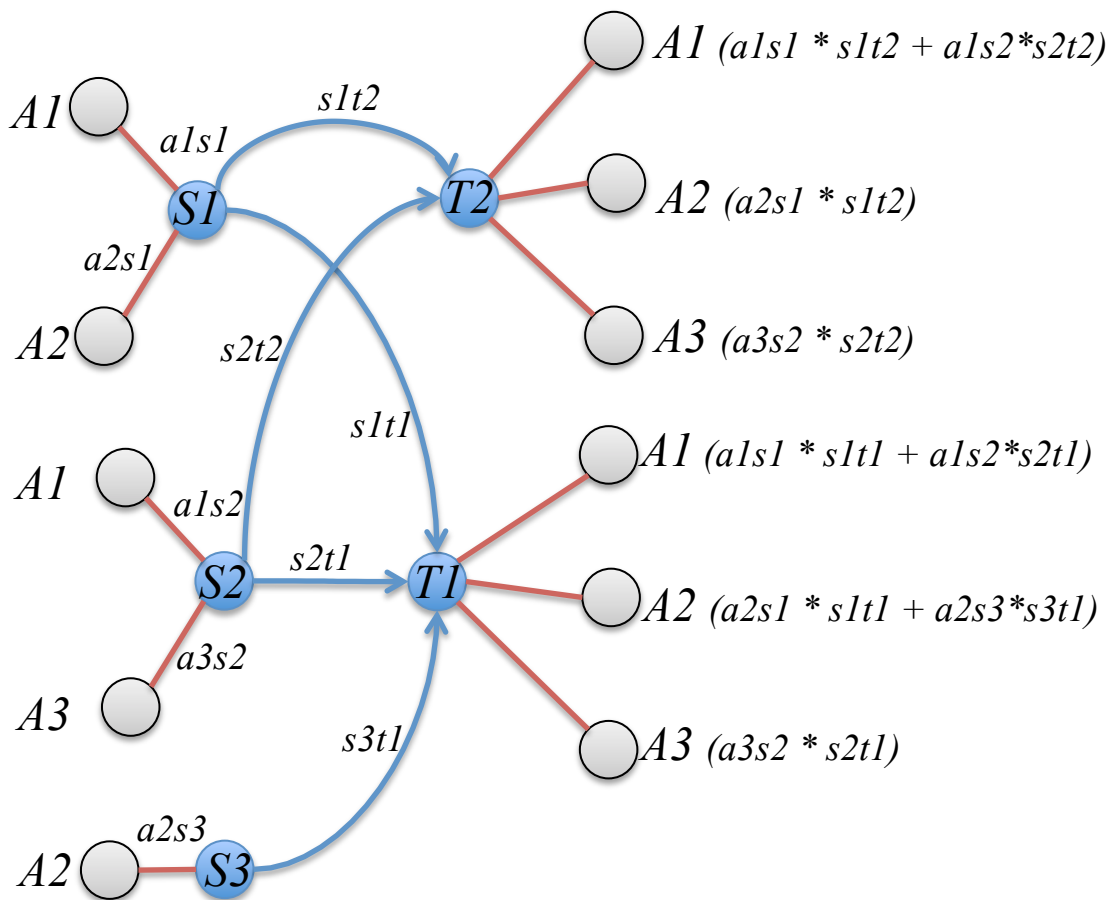
This translates as matrix multiplication of Annotation Confidence Matrix and Orthofuzz matrix

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

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

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

Cluster by Random Walking

Data:

Shared among threads:

Transferred to threads

Random Walking with Restart

Create population of Scores

Estimate Statistics
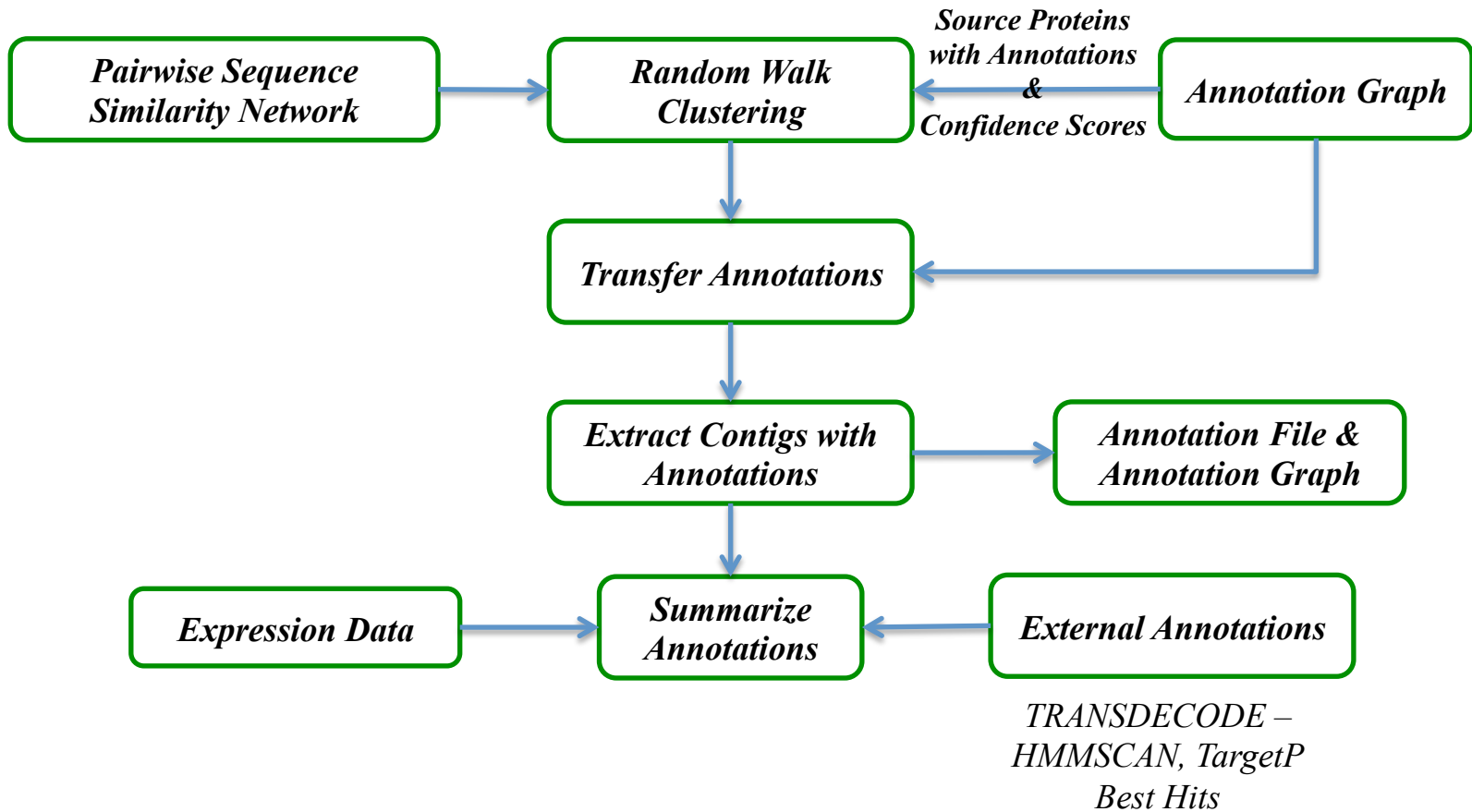
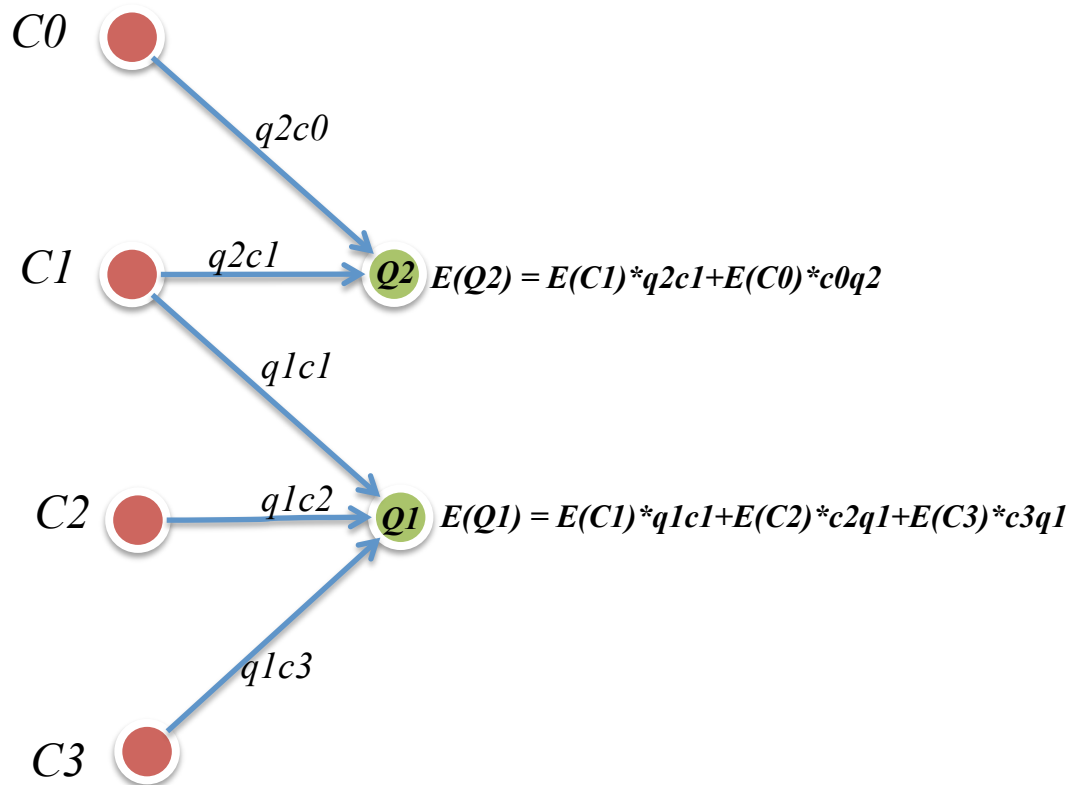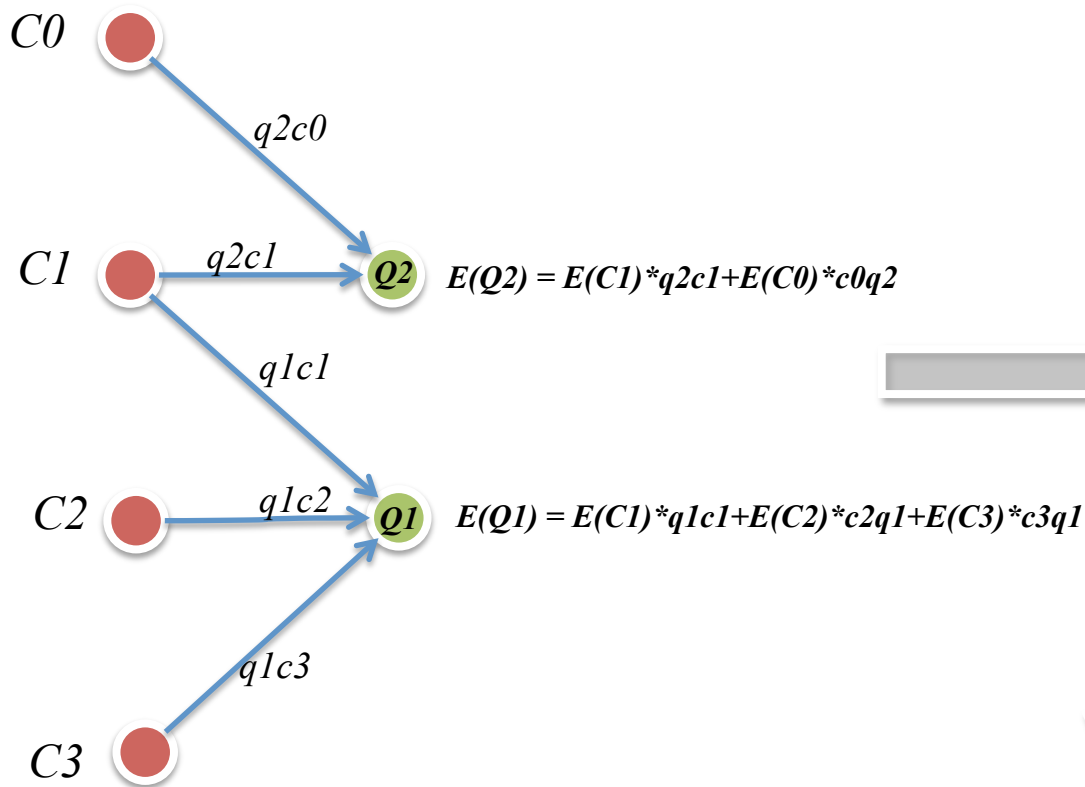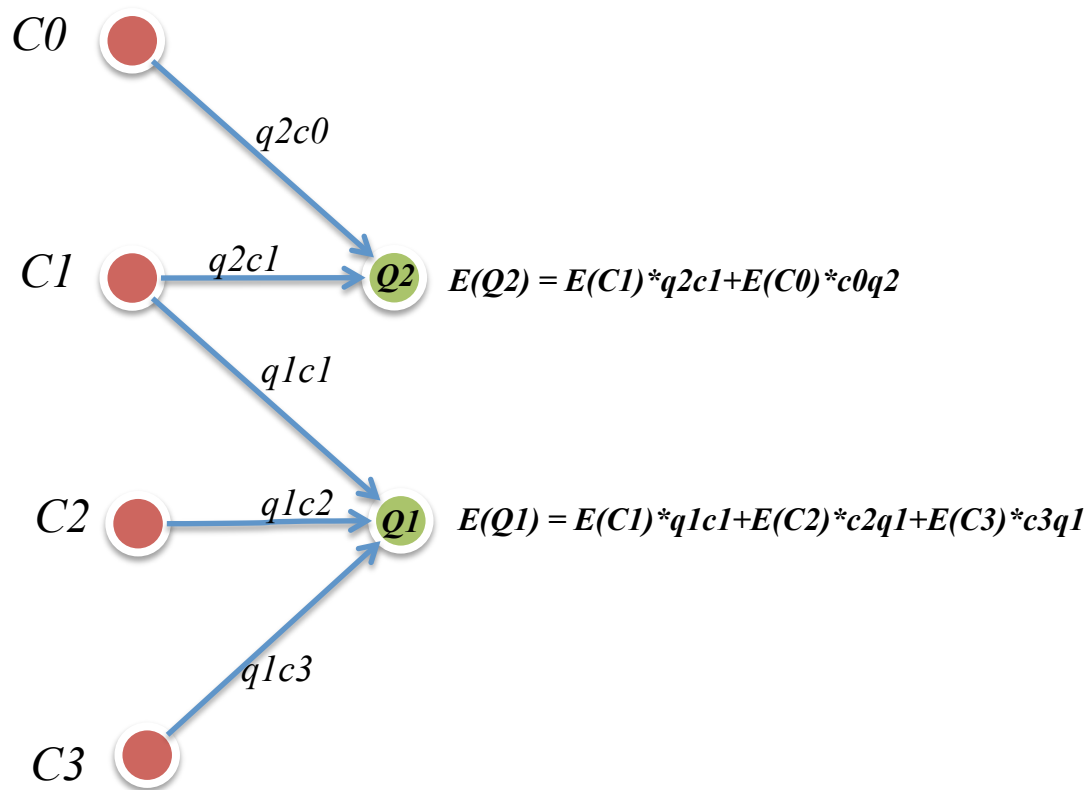*Figure 1: Protocol used for assigning functional annotations to the de-novo assembled contigs*

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. qxcy is the within species normalized orthofuzzscore (qx,cy) obtained by querying the network using the query set Q.

$C0$

$C1$

$q2c0$

$q2c1$ $Q2$ $E(Q2) = E(C1)*q2c1+E(C0)*c0q2$

$q1c1$

$C2$ $q1c2$ $Q1$ $E(Q1) = E(C1)*q1c1+E(C2)*c2q1+E(C3)*c3q1$

$q1c3$

$C3$

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

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

C0

q2c0

C1   q2c1   Q2 $E(Q2) = E(C1)*q2c1+E(C0)*c0q2$

q1c1

C2   q1c2   Q1 $E(Q1) = E(C1)*q1c1+E(C2)*c2q1+E(C3)*c3q1$

q1c3

C3

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

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

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

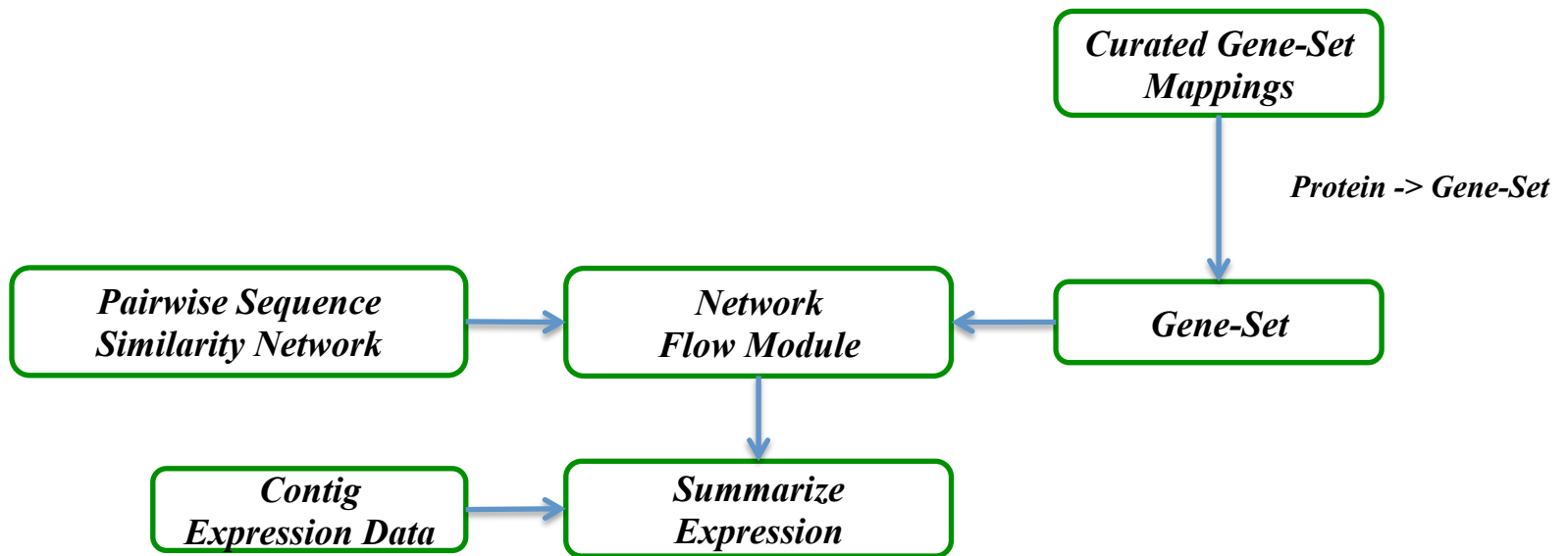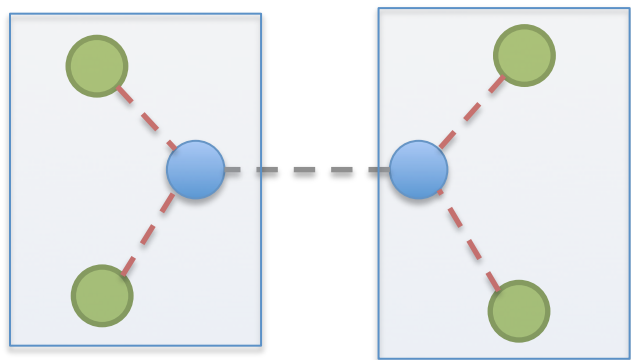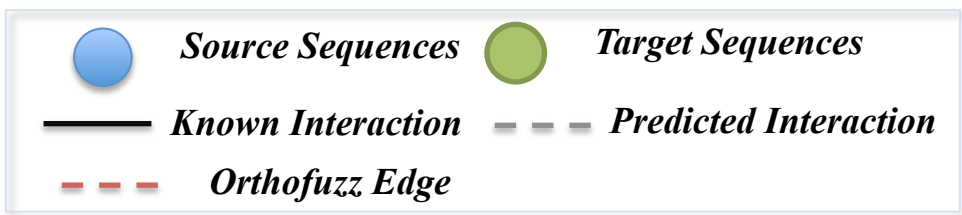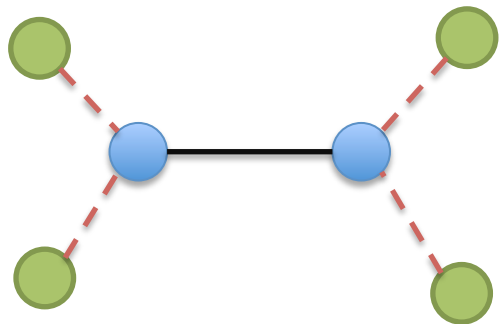| Symbol	Description
Q	*Set of Query Sets*
N_Q	*Total Number of Query Sets*
N_C	*Total Number of Expressed Contigs*
QC	*Query Contig Orthofuzz Matrix*
$qxcy$	*Species.normalized orthofuzz score of qx to cy*
Ce	*Vector containing expression values of Contigs*
e_{cy}	*Expression value of contig y*
Qe	*Vector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

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

Source Sequences ⬤   **Target Sequences** ⬤

—— **Known Interaction**   - - - **Predicted Interaction**

- - - **Orthofuzz Edge**

Expression Levels

	Conditions...		
Interaction Unit 1			
Interaction Unit 2			

Are there interaction units whose expression levels vary significantly?
Are the predicted units co-expressed ?

Is there evidence for protein interaction clusters ?

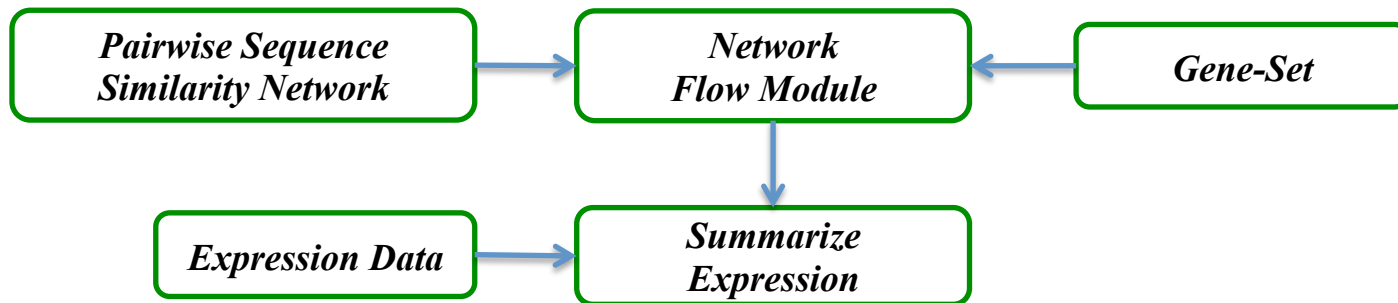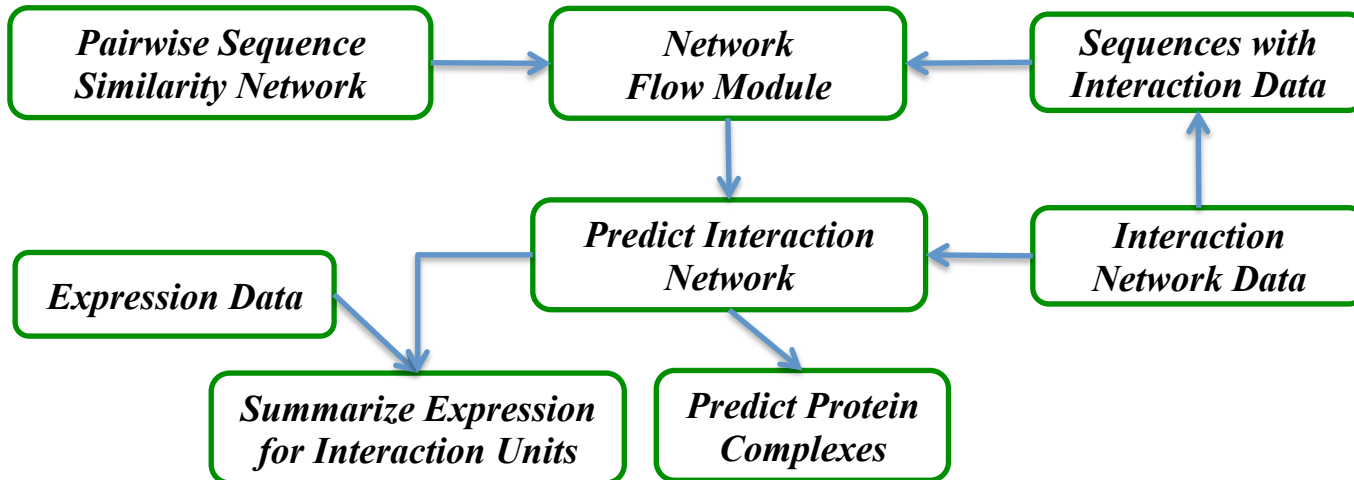*Figure: Protocol used for estimating expression levels of gene-sets*

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