EasyVectorOmics

Tensor Omics Team

May 23, 2025

1 Overview of EasyVectorOmics

The EasyVectorOmics prototype implements a streamlined pipeline for analyzing gene expression data and evolutionary relationships between genes. The pipeline integrates tools for data normalization, phylogenetic classification, synteny analysis, and geometric analysis of gene expression vectors. This document outlines the current implementation details and provides the mathematical formulation of the algorithms used in the prototype.

2 Pipeline Steps

2.1 Data Normalization

Gene expression data is normalized to ensure comparability across samples. The normalization process includes:

1. Standard Deviation Normalization: Each gene vector $\mathbf{x}_i$ is scaled by its standard deviation magnitude, estimated as:

$$\sigma_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n} x_{ij}^2} \tag{1}$$

   where x_{ij} is the expression of gene i in tissue j, and n is the number of tissues.

   The normalized expression vector $\tilde{\mathbf{x}}_i$ is then:

$$\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\sigma_i} \tag{2}$$

   If $\sigma_i = 0$, a fallback value of 1 is used to avoid division by zero.

2. Quantile normalization across samples to reduce global expression biases.

3. Log-transformation: $x \mapsto \log(1 + x)$ to stabilize variance and suppress outliers.

4. For stress response studies, log fold-changes are computed:

$$\Delta_{\text{stress}} = \log(1 + x_{\text{stress}}) - \log(1 + x_{\text{control}})$$

   This effectively makes each stress related axis show the fold change in gene expression under stress, e.g. "drought in root divided by control root".

2.2 Protein Similarity Calculation

Protein similarity is calculated using the harmonic mean of sequence identity (p_{ident}) and overlap percentage (o_{lap}) between protein pairs. The overlap is computed as:

$$o_{\text{lap}} = \frac{(\text{qend} - \text{qstart}) + (\text{send} - \text{sstart})}{\text{qlen} + \text{slen}} \cdot 100$$

where:

- $\mathrm{qstart}, \mathrm{qend}$: Start and end positions of the query sequence.
- $\mathrm{sstart}, \mathrm{send}$: Start and end positions of the subject sequence.
- $\mathrm{qlen}, \mathrm{slen}$: Lengths of the query and subject sequences.

The harmonic mean is then calculated as:

$$\text{Harmonic Mean Similarity} = \frac{2 \cdot p_{\text{ident}} \cdot o_{\text{lap}}}{p_{\text{ident}} + o_{\text{lap}}}$$

This ensures that the similarity measure is symmetric and penalizes discrepancies between sequence identity and overlap.

2.3 Phylogenetic Classification of Genes

Phylogenetic classification is performed using resolved gene trees and orthogroup data. The algorithm classifies genes into the following categories:

- **Conserved Orthologs:** Genes with a one-to-one relationship across species, identified as the best match between subtrees at speciation nodes.
- **Inparalogs:** Genes duplicated after the last common ancestor of the species being compared.
- **Outparalogs:** Genes duplicated before the last common ancestor.
- **Source-copy Inparalogs:** Pairs consisting of a conserved protein and a non-conserved protein from the same species.
- **Source-copy Orthologs:** Pairs consisting of a conserved protein and a non-conserved protein from different species.

2.3.1 Algorithm Description (Tree-Based Classification)

The classification algorithm processes resolved gene trees as follows:

1. **Input Data:** The algorithm requires Newick tree files, a ‘Duplications.tsv‘ file indicating duplication nodes, an ‘Orthogroups.tsv‘ file mapping proteins to orthogroups, and a score dictionary (pickle file) containing the best match scores between genes.
2. **Starting at the Root:** The algorithm begins analyzing the tree from the root node.
3. **Speciation and Subtree Comparison:** At each speciation node, the leaves of the left and right subtrees are compared. Protein IDs and their corresponding species are taken into account.
4. **Best Pair Identification:** For each species, the algorithm identifies the best gene pair between the left and right subtrees using the score dictionary.
5. **Ortholog Classification:** The identified best pairs are classified as conserved orthologs.
6. **Inparalog and Outparalog Identification:** Genes not classified as conserved orthologs are further analyzed to determine whether they are inparalogs or outparalogs based on their duplication history.
7. **Source-copy cases:** The algorithm identifies:
    - **Source-copy Inparalogs:** Pairs consisting of a conserved protein and a non-conserved protein from the same species.
    - **Source-copy Orthologs:** Pairs consisting of a conserved protein and a non-conserved protein from different species.
8. **Handling Small Families:** For orthogroups with fewer than 4 genes (where OrthoFinder does not generate trees), the ‘tree_rest.py‘ script classifies genes into orthologs, inparalogs, and source-copy cases.

2.3.2 Output

The results are written into various output files containing information about:

- Conserved Orthologs
- Inparalogs
- Outparalogs
- Source-copy Inparalogs
- Source-copy Orthologs

2.4 Synteny-Based Gene Classification

The synteny analysis algorithm identifies gene relationships based on genomic neighborhoods and shared genes between species. The process is as follows:

1. **Input Data:** The algorithm requires:
    - GTF files for each species, containing gene coordinates and genomic features.
    - Tandem gene information, which groups genes located close to each other on the genome.
    - BLAST results, providing pairwise relationships between genes.
2. **Gene Neighborhood Construction:**
    - Genes are grouped by chromosome and sorted by their genomic positions.
    - For each gene, a neighborhood is constructed by selecting a fixed number of neighboring genes on both sides (default: 10 neighbors).
    - Tandem genes are handled by selecting only the representative gene from each tandem group to avoid redundancy.
3. **Relationship Dictionary:**
    - A dictionary is created from the BLAST results, where each gene is mapped to its related genes along with their percent identity scores.
4. **Shared Gene Analysis:**
    - For each gene, its neighborhood is compared with the neighborhoods of related genes from other species.
    - The algorithm uses the **Hopcroft-Karp algorithm** to find the maximum bipartite matching between the two neighborhoods. This ensures that the maximum number of unique gene pairs is identified between the neighborhoods.
5. **Output:**
    - The results are saved in a file containing pairs of genes from different species, along with the count of shared genes between their neighborhoods.

2.4.1 Hopcroft-Karp Algorithm

The Hopcroft-Karp algorithm is a graph-theoretic method used to find the maximum matching in a bipartite graph. In this context:

- Each neighborhood is treated as a set of nodes in a bipartite graph.
- Edges are added between nodes (genes) if they are related based on the BLAST relationship dictionary.
- The algorithm identifies the maximum number of unique gene pairs (matching) between the two neighborhoods.

This approach ensures that the synteny analysis captures the most significant relationships between genes in different species, while avoiding redundancy caused by tandem genes or overlapping neighborhoods.

2.5 Centroid Calculation for Expression Vectors

The centroid of a gene family is calculated as the mean of the expression vectors of genes marked as orthologs:

$$\vec{o} = \frac{1}{n}\sum_{i=1}^{n}\vec{v}_i$$

where $\vec{v}_i$ is the expression vector of the i-th ortholog in the family.

2.6 Euclidean Distance Calculation

The Euclidean distance calculation step identifies genes with significant deviations in expression patterns by comparing their expression vectors to the centroid of their respective orthogroups. This step also computes pairwise distances between genes within orthogroups for further analysis.

1. **Input Data:**
   - Normalized expression data for all genes.
   - Centroid vectors for each orthogroup.
   - Orthogroup mappings for genes.
   - Ortholog and paralog relationships.
2. **Distance to Centroid:**
   - For each gene, the Euclidean distance to the centroid of its orthogroup is calculated:

     $$d_i = \sqrt{\sum_{j=1}^{n}(x_{ij} - o_j)^2}$$

     where x_{ij} is the expression value of gene i in tissue j, and o_j is the centroid value for tissue j.
   - The distances are normalized by dividing by the maximum pairwise distance between orthologs within the same orthogroup, resulting in the Relative Divergence Index (RDI):

     $$\mathrm{RDI}_i = \frac{d_i}{\max(d_{ij})}$$

     where $\max(d_{ij})$ is the maximum pairwise distance between orthologs in the orthogroup.
3. **Outlier Identification:**
   - Genes with an RDI above the 95th percentile are flagged as outliers.

- A second level of filtering identifies "outliers of outliers," defined as the top 5% of outliers with the highest RDI values.

4. **Pairwise Distances:**

   - Pairwise Euclidean distances are computed between all gene pairs within orthogroups, including orthologs and paralogs.
   - These distances are used to analyze relationships between genes and to calculate the maximum pairwise distance for normalizing the RDI.

5. **Purpose of Outlier Identification:**

   - Outliers represent genes with significant deviations in expression patterns, which may indicate biological relevance, such as involvement in specific pathways or responses to experimental conditions.
   - Only outliers are considered in subsequent analyses to focus on genes with the most significant expression changes.

2.7 Evolutionary Angle Calculations

The evolutionary angle calculations provide insights into the geometric relationships between gene expression vectors and their centroids, as well as their alignment with the space diagonal. These calculations are used to analyze tissue versatility, divergence, and adaptation of gene families.

1. **Input Data:**

   - Normalized expression vectors for genes.
   - Centroid vectors for orthogroups.
   - Pairwise relationships between genes (orthologs and paralogs).

2. **Tissue Versatility Calculation:**

   - The angle between a gene's expression vector $\vec{v}_i$ and the space diagonal $\vec{d}$ (a vector with equal components in all dimensions) is calculated using the cosine similarity:

   $$\cos(\theta) = \frac{\vec{v}_i \cdot \vec{d}}{\|\vec{v}_i\| \cdot \|\vec{d}\|}$$

   where $\vec{d} = (1, 1, \ldots, 1)/\sqrt{n}$, with n being the number of dimensions (e.g., tissues or conditions).
   - A smaller angle (θ) indicates higher tissue versatility, meaning the gene is uniformly expressed across tissues. Conversely, a larger angle indicates tissue specificity, where the gene's expression is concentrated in a subset of tissues.

3. **Clock Plot Projection and Rotation Angle Computation:**

   - Gene expression vectors and their centroids are projected into the **Relative Axes Plane (RAP)**, which is orthogonal to the space diagonal and intersects the origin.
   - The projection is used to compute the rotation angle (ϕ) between the centroid and the gene's expression vector in the RAP:

   $$\phi = \arccos\left(\frac{\vec{p}_{\text{proj}} \cdot \vec{o}_{\text{proj}}}{\|\vec{p}_{\text{proj}}\| \cdot \|\vec{o}_{\text{proj}}\|}\right)$$

   where $\vec{p}_{\text{proj}}$ and $\vec{o}_{\text{proj}}$ are the projections of the gene vector and centroid vector, respectively, onto the RAP.

- This angle quantifies the divergence of a gene's expression pattern from its family centroid in the RAP, providing insights into functional differentiation and adaptation.

4. **Pairwise Angle Calculations:**

   - The angle between the orthogonal components of two gene vectors (e.g., orthologs or paralogs) is calculated to analyze their relative divergence:

   $$\theta_{\text{pairwise}} = \arccos\left(\frac{\vec{v}_1 \cdot \vec{v}_2}{\|\vec{v}_1\| \cdot \|\vec{v}_2\|}\right)$$

   - These pairwise angles are used to study evolutionary relationships and functional shifts within orthogroups.

5. **Purpose of Angle Calculations:**

   - **Tissue Versatility:** Identify genes with uniform versus tissue-specific expression patterns.
   - **Functional Divergence:** Quantify the divergence of gene expression patterns from their family centroid.
   - **Evolutionary Adaptation:** Analyze the relative shifts in expression patterns between orthologs and paralogs.

2.8 Functional Annotation and Word Cloud Generation

The functional annotation and word cloud generation step provides a visual summary of the biological functions associated with outlier genes. This step relies on annotations generated using Prot-Scriber and focuses on identifying enriched terms in the descriptions of outlier genes.

1. **Input Data:**

   - Outlier gene table, including gene IDs and their classification as outliers.
   - Functional annotations generated by Prot-Scriber, which assigns short human-readable descriptions (HRDs) to query biological sequences based on sequence similarity search results (e.g., BLAST or DIAMOND).

2. **Prot-Scriber Annotations:**

   - Prot-Scriber assigns HRDs to query sequences by performing lexical analysis on the descriptions of BLAST hits.
   - These HRDs provide concise descriptions of the biological functions or roles of the sequences, enabling downstream analysis of gene families or individual genes.

3. **Word Cloud Generation:**

   - The script filters the annotations to include only those corresponding to outlier genes.
   - A text corpus is built from the HRDs of outlier genes, and preprocessing steps such as lowercasing, punctuation removal, and whitespace stripping are applied.
   - A term-document matrix is created to compute word frequencies, and overly frequent terms (appearing in more than 80% of descriptions) are excluded.
   - A word cloud is generated to visualize the most frequent terms associated with outlier genes, providing insights into their biological relevance.

4. **Statistical Enrichment Analysis:**

   - A Fisher's exact test is applied to compare the presence of terms in outlier versus non-outlier genes.

- The results include p-values and odds ratios for each term, highlighting terms significantly enriched in outlier genes.

5. **Output:**

    - A word cloud image summarizing the most frequent terms in outlier gene annotations.
    - A table of word frequencies and statistical enrichment results, saved as TSV files for further analysis.

6. **Purpose:**

    - The word cloud provides an intuitive visualization of the biological functions associated with outlier genes.
    - The statistical enrichment analysis identifies terms that are significantly overrepresented in outlier genes, offering insights into their potential roles in specific pathways or conditions.