# Appendix

# Appendix

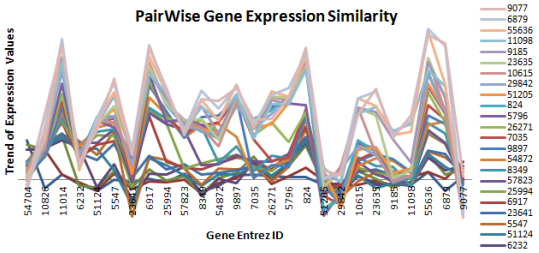
## Experimental Results of the Cluster Analysis of Breast Cancer dataset, Lymphoma dataset and Central Nervous System dataset

The results for the experiments involving semantic similarity, sequence similarity, internal validity measures, stability measures and biological measures on the (a) Breast Cancer dataset (b) Lymphoma dataset and (c) Embryonal Tumours of the Central Nervous System (CNS) dataset are given here.

### A. The Pair-wise Gene Expression Similarity Matrix

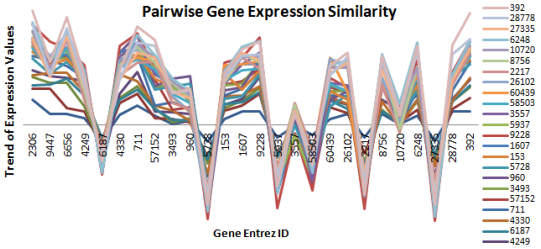
The pair-wise gene expression similarity is calculated using Pearson Correlation for (a) Breast Cancer dataset, (b) Lymphoma dataset and (c) Embryonal Tumours of the Central Nervous System (CNS) dataset. The similarity matrix value for some of the genes is shown in *Figure* A-1 below along with the plots of the values. These expression values will be used subsequently for comparison with the semantic similarity values for the corresponding pair of genes in Section C.

| Entrez ID | 54704 | 10826 | 11014 | 6232 | 51124 | ... | 8573 | 1112 | 1644 | 29089 | 10112 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **54704** | 1.0000 | -0.2318 | 0.1263 | -0.1226 | -0.3090 | ... | -0.1381 | 0.1940 | 0.0089 | -0.1015 | 0.0675 |
| **10826** | -0.2318 | 1.0000 | -0.0121 | 0.0480 | 0.0434 | ... | -0.0545 | 0.3128 | 0.0828 | -0.2668 | -0.3253 |
| **11014** | 0.1263 | -0.0121 | 1.0000 | -0.0989 | 0.0617 | ... | 0.0734 | -0.1616 | 0.0045 | 0.3310 | 0.3715 |
| **6232** | -0.1226 | 0.0480 | -0.0989 | 1.0000 | -0.0053 | ... | -0.0128 | 0.0442 | -0.1371 | 0.0138 | 0.0367 |
| **51124** | -0.3090 | 0.0434 | 0.0617 | -0.0053 | 1.0000 | ... | 0.1252 | -0.2141 | 0.0347 | 0.3105 | 0.0984 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **8573** | -0.1381 | -0.0545 | 0.0734 | -0.0128 | 0.1252 | ... | 1.0000 | -0.2261 | -0.0370 | 0.0799 | 0.0547 |
| **1112** | 0.1940 | 0.3128 | -0.1616 | 0.0442 | -0.2141 | ... | -0.2261 | 1.0000 | 0.0950 | -0.3624 | -0.3216 |
| **1644** | 0.0089 | 0.0828 | 0.0045 | -0.1371 | 0.0347 | ... | -0.0370 | 0.0950 | 1.0000 | 0.2910 | 0.1394 |
| **29089** | -0.1015 | -0.2668 | 0.3310 | 0.0138 | 0.3105 | ... | 0.0799 | -0.3624 | 0.2910 | 1.0000 | 0.7255 |
| **10112** | 0.0675 | -0.3253 | 0.3715 | 0.0367 | 0.0984 | ... | 0.0547 | -0.3216 | 0.1394 | 0.7255 | 1.0000 |



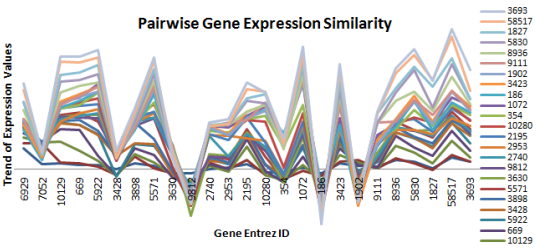

(a) Breast Cancer dataset

| Entrez ID | 2306 | 9447 | 6656 | 4249 | 6187 | ... | 5118 | 8658 | 8339 | 1638 | 27251 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **2306** | 1.0000 | 0.4233 | 0.4474 | 0.2661 | -0.5872 | ... | 0.0738 | 0.5118 | -0.1986 | 0.5335 | 0.5779 |
| **9447** | 0.4233 | 1.0000 | 0.2137 | 0.2733 | -0.3355 | ... | -0.2359 | 0.4966 | -0.0014 | 0.508 | 0.462 |
| **6656** | 0.4474 | 0.2137 | 1.0000 | 0.1868 | -0.0316 | ... | -0.0681 | 0.1283 | 0.1962 | 0.0835 | 0.0476 |
| **4249** | 0.2661 | 0.2733 | 0.1868 | 1.0000 | -0.5813 | ... | -0.3035 | 0.3561 | -0.1154 | 0.4096 | 0.3661 |
| **6187** | -0.587 | -0.336 | -0.032 | -0.581 | 1.0000 | ... | -0.1433 | -0.3073 | 0.0909 | -0.61 | -0.5206 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **5118** | 0.0738 | -0.236 | -0.068 | -0.303 | -0.1433 | ... | 1.0000 | -0.2273 | 0.0851 | 0.0219 | 0.1727 |
| **8658** | 0.5118 | 0.4966 | 0.1283 | 0.3561 | -0.3073 | ... | -0.2273 | 1.0000 | -0.1802 | 0.4181 | 0.4698 |
| **8339** | -0.199 | -0.001 | 0.1962 | -0.115 | 0.09091 | ... | 0.0851 | -0.1802 | 1.0000 | -0.059 | -0.149 |
| **1638** | 0.5335 | 0.508 | 0.0835 | 0.4096 | -0.6101 | ... | 0.0219 | 0.4181 | -0.0591 | 1.0000 | 0.7542 |
| **27251** | 0.5779 | 0.462 | 0.0476 | 0.3661 | -0.5206 | ... | 0.1727 | 0.4698 | -0.149 | 0.7542 | 1.0000 |




(b) Lymphoma dataset

| Entrez ID | 6929 | 7052 | 10129 | 669 | 5922 | ... | 10924 | 1609 | 5108 | 3796 | 273 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6929 | 1.0000 | 0.2175 | 0.2882 | 0.2098 | 0.2024 | ... | 0.2325 | -0.014 | 0.2793 | -0.0415 | 0.01508 |
| 7052 | 0.2175 | 1.0000 | 0.0275 | 0.0743 | -0.091 | ... | 0.046 | -0.566 | 0.2279 | 0.0452 | -0.0232 |
| 10129 | 0.2882 | 0.0275 | 1.0000 | 0.5895 | 0.291 | ... | 0.3857 | 0.1477 | 0.3121 | 0.1576 | 0.25499 |
| 669 | 0.2098 | 0.0743 | 0.5895 | 1.0000 | 0.2392 | ... | 0.576 | 0.2332 | 0.3079 | -0.0608 | 0.2204 |
| 5922 | 0.2024 | -0.091 | 0.291 | 0.2392 | 1.0000 | ... | 0.3203 | 0.3052 | 0.4333 | -0.099 | -0.056 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 10924 | 0.2325 | 0.046 | 0.3857 | 0.576 | 0.3203 | ... | 1.0000 | 0.0395 | 0.3365 | -0.0595 | -0.0393 |
| 1609 | -0.014 | -0.566 | 0.1477 | 0.2332 | 0.3052 | ... | 0.0395 | 1.0000 | 0.2063 | 0.1651 | 0.33877 |
| 5108 | 0.2793 | 0.2279 | 0.3121 | 0.3079 | 0.4333 | ... | 0.3365 | 0.2063 | 1.0000 | 0.0752 | 0.13313 |
| 3796 | -0.041 | 0.0452 | 0.1576 | -0.061 | -0.099 | ... | -0.059 | 0.1651 | 0.0752 | 1.0000 | 0.13604 |
| 273 | 0.0151 | -0.023 | 0.255 | 0.2204 | -0.056 | ... | -0.039 | 0.3388 | 0.1331 | 0.136 | 1.0000 |



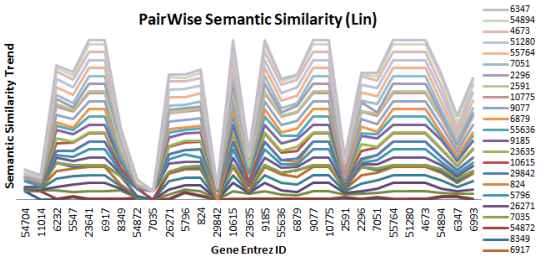

(c) Embryonal Tumours of the Central Nervous System

**Figure A-1:** Pair-Wise Gene Expression Similarity Matrix for (a) Breast Cancer dataset, (b) Lymphoma dataset and (c) Embryonal Tumours of the Central Nervous System (CNS)

## B. The Pair-wise Semantic Similarity Matrix

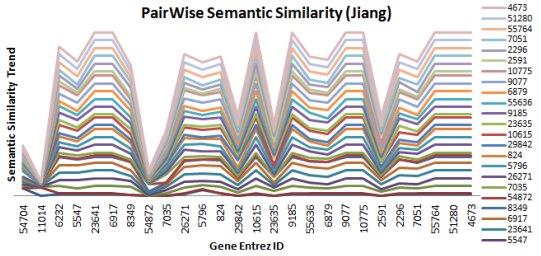
The pair-wise semantic similarity matrix for the Lin, Jiang and Conrath and Wang measures for the Breast Cancer dataset are calculated and the semantic similarity values and plots for some of the genes is shown in *Figure* A-2 below:

| Entrez ID | 54704 | 11014 | 6232 | 5547 | 23641 | ... | 8573 | 1112 | 1644 | 29089 | 10112 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 54704 | 1.000 | 0.000 | 0.000 | 0.320 | 0.000 | ... | 0.218 | 0.000 | 0.196 | 0.279 | 0.000 |
| 11014 | 0.000 | 1.069 | 0.089 | 0.089 | 0.089 | ... | 0.089 | 0.089 | 0.089 | 0.066 | 0.068 |
| 6232 | 0.000 | 0.089 | 1.024 | 0.498 | 1.000 | ... | 0.498 | 0.558 | 0.498 | 0.183 | 0.213 |
| 5547 | 0.320 | 0.089 | 0.498 | 1.000 | 1.000 | ... | 0.594 | 0.382 | 0.585 | 0.262 | 0.201 |
| 23641 | 0.000 | 0.089 | 1.000 | 1.000 | 1.000 | ... | 1.000 | 1.000 | 1.000 | 0.346 | 0.402 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 8573 | 0.218 | 0.089 | 0.498 | 0.594 | 1.000 | ... | 1.000 | 0.382 | 0.589 | 0.268 | 0.201 |
| 1112 | 0.000 | 0.089 | 0.558 | 0.382 | 1.000 | ... | 0.382 | 1.000 | 0.382 | 0.164 | 0.185 |
| 1644 | 0.196 | 0.089 | 0.498 | 0.585 | 1.000 | ... | 0.589 | 0.382 | 1.000 | 0.256 | 0.201 |
| 29089 | 0.279 | 0.066 | 0.183 | 0.262 | 0.346 | ... | 0.268 | 0.164 | 0.256 | 1.000 | 0.290 |
| 10112 | 0.000 | 0.068 | 0.213 | 0.201 | 0.402 | ... | 0.201 | 0.185 | 0.201 | 0.290 | 1.000 |



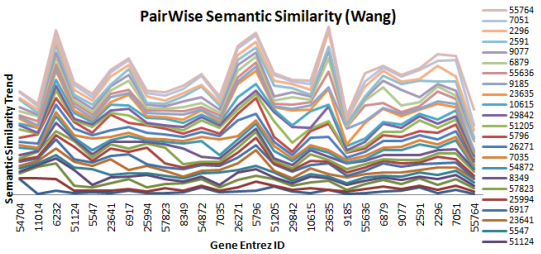

(a) Lin Semantic Similarity

| Entrez ID | 54704 | 11014 | 6232 | 5547 | 23641 | ... | 8573 | 1112 | 1644 | 29089 | 10112 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 54704 | 1.000 | 0.000 | 0.267 | 0.267 | 0.267 | ... | 0.267 | 0.267 | 0.267 | 0.212 | 0.085 |
| 11014 | 0.000 | 1.130 | 0.045 | 0.045 | 0.045 | ... | 0.045 | 0.045 | 0.045 | 0.000 | 0.000 |
| 6232 | 0.267 | 0.045 | 1.012 | 0.692 | 1.000 | ... | 0.707 | 0.722 | 0.684 | 0.457 | 0.559 |
| 5547 | 0.267 | 0.045 | 0.692 | 1.000 | 1.000 | ... | 0.537 | 0.559 | 0.509 | 0.412 | 0.481 |
| 23641 | 0.267 | 0.045 | 1.000 | 1.000 | 1.000 | ... | 1.000 | 1.000 | 1.000 | 0.574 | 0.665 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 8573 | 0.267 | 0.045 | 0.707 | 0.537 | 1.000 | ... | 1.000 | 0.568 | 0.528 | 0.423 | 0.481 |
| 1112 | 0.267 | 0.045 | 0.722 | 0.559 | 1.000 | ... | 0.568 | 1.000 | 0.554 | 0.286 | 0.369 |
| 1644 | 0.267 | 0.045 | 0.684 | 0.509 | 1.000 | ... | 0.528 | 0.554 | 1.000 | 0.412 | 0.481 |
| 29089 | 0.212 | 0.000 | 0.457 | 0.412 | 0.574 | ... | 0.423 | 0.286 | 0.412 | 1.000 | 0.409 |
| 10112 | 0.085 | 0.000 | 0.559 | 0.481 | 0.665 | ... | 0.481 | 0.369 | 0.481 | 0.409 | 1.000 |




(b) Jiang and Conrath Semantic Similarity

| Entrez ID | 54704 | 11014 | 6232 | 51124 | 5547 | ... | 8573 | 1112 | 1644 | 29089 | 10112 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 54704 | 1.000 | 0.081 | 0.301 | 0.121 | 0.182 | ... | 0.133 | 0.178 | 0.472 | 0.411 | 0.138 |
| 11014 | 0.081 | 1.000 | 0.691 | 0.151 | 0.097 | ... | 0.280 | 0.163 | 0.130 | 0.199 | 0.327 |
| 6232 | 0.301 | 0.691 | 1.000 | 0.308 | 0.225 | ... | 0.260 | 0.374 | 0.327 | 0.484 | 0.399 |
| 51124 | 0.121 | 0.151 | 0.308 | 1.000 | 0.148 | ... | 0.347 | 0.258 | 0.157 | 0.245 | 0.345 |
| 5547 | 0.182 | 0.097 | 0.225 | 0.148 | 1.000 | ... | 0.215 | 0.223 | 0.130 | 0.277 | 0.143 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 8573 | 0.133 | 0.280 | 0.260 | 0.347 | 0.215 | ... | 1.000 | 0.278 | 0.153 | 0.215 | 0.361 |
| 1112 | 0.178 | 0.163 | 0.374 | 0.258 | 0.223 | ... | 0.278 | 1.000 | 0.225 | 0.245 | 0.305 |
| 1644 | 0.472 | 0.130 | 0.327 | 0.157 | 0.130 | ... | 0.153 | 0.225 | 1.000 | 0.298 | 0.183 |
| 29089 | 0.411 | 0.199 | 0.484 | 0.245 | 0.277 | ... | 0.215 | 0.245 | 0.298 | 1.000 | 0.225 |
| 10112 | 0.138 | 0.327 | 0.399 | 0.345 | 0.143 | ... | 0.361 | 0.305 | 0.183 | 0.225 | 1.000 |

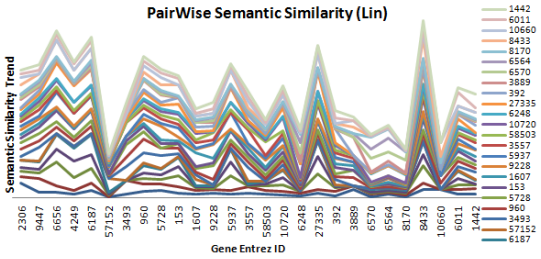



(c) Wang Semantic Similarity

**Figure A-2:** Pair-Wise Semantic Similarity Matrix using (a) Lin, (b) Jiang and Conrath and (c) Wang for Breast Cancer dataset

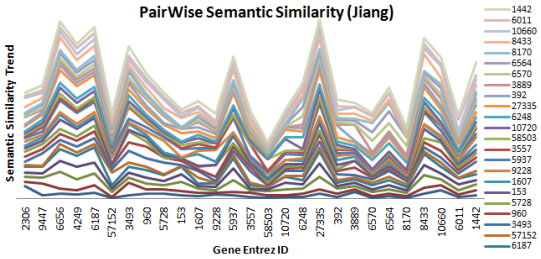
The pair-wise semantic similarity matrix given in *Figure* A-3 shows the values for Lin, Jiang and Conrath and Wang measures for the Lymphoma dataset along with the plots.

| Entrez ID | 2306 | 9447 | 6656 | 4249 | 6187 | ... | 8658 | 7157 | 2747 | 5118 | 1638 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2306 | 1.078 | 0.428 | 0.407 | 0.287 | 0.504 | ... | 0.45 | 0.534 | 0.178 | 0.452 | 0.485 |
| 9447 | 0.428 | 0.944 | 0.423 | 0.263 | 0.555 | ... | 0.544 | 0.541 | 0.174 | 0.301 | 0.218 |
| 6656 | 0.407 | 0.423 | 1.73 | 0.923 | 1.07 | ... | 0.756 | 0.481 | 0.527 | 0.236 | 0.555 |
| 4249 | 0.287 | 0.263 | 0.923 | 1.143 | 0.979 | ... | 0.597 | 0.292 | 0.616 | 0.165 | 0.437 |
| 6187 | 0.504 | 0.555 | 1.07 | 0.979 | 1.432 | ... | 0.783 | 0.513 | 0.549 | 0.151 | 0.485 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 8658 | 0.45 | 0.544 | 0.756 | 0.597 | 0.783 | ... | 1.037 | 0.557 | 0.582 | 0.186 | 0.359 |
| 7157 | 0.534 | 0.541 | 0.481 | 0.292 | 0.513 | ... | 0.557 | 1.106 | 0.179 | 0.218 | 0.23 |
| 2747 | 0.178 | 0.174 | 0.527 | 0.616 | 0.549 | ... | 0.382 | 0.179 | 1.066 | 0.138 | 0.355 |
| 5118 | 0.452 | 0.301 | 0.236 | 0.165 | 0.151 | ... | 0.186 | 0.218 | 0.138 | 1 | 0.332 |
| 1638 | 0.485 | 0.218 | 0.555 | 0.437 | 0.485 | ... | 0.359 | 0.23 | 0.355 | 0.332 | Inf |



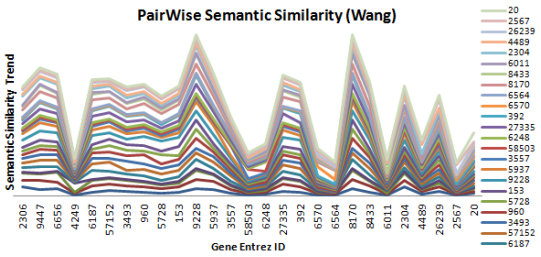

(a) Lin Semantic Similarity

| Entrez ID | 2306 | 9447 | 6656 | 4249 | 6187 | ... | 8658 | 7157 | 2747 | 5118 | 1638 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2306 | 1.077 | 0.325 | 0.412 | 0.36 | 0.533 | ... | 0.374 | 0.467 | 0.173 | 0.503 | 0.409 |
| 9447 | 0.325 | 0.946 | 0.471 | 0.351 | 0.589 | ... | 0.412 | 0.514 | 0.156 | 0.381 | 0.179 |
| 6656 | 0.412 | 0.471 | 1.387 | 0.914 | 0.977 | ... | 0.707 | 0.504 | 0.578 | 0.521 | 0.638 |
| 4249 | 0.36 | 0.351 | 0.914 | 1.065 | 0.931 | ... | 0.595 | 0.395 | 0.608 | 0.389 | 0.327 |
| 6187 | 0.533 | 0.589 | 0.977 | 0.931 | 1.254 | ... | 0.761 | 0.56 | 0.541 | 0.403 | 0.405 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 8658 | 0.374 | 0.412 | 0.707 | 0.595 | 0.761 | ... | 1.032 | 0.467 | 0.263 | 0.176 | 0.182 |
| 7157 | 0.467 | 0.514 | 0.504 | 0.395 | 0.56 | ... | 0.467 | 1.075 | 0.176 | 0.284 | 0.193 |
| 2747 | 0.173 | 0.156 | 0.578 | 0.608 | 0.541 | ... | 0.263 | 0.176 | 1.107 | 0.197 | 0.132 |
| 5118 | 0.503 | 0.381 | 0.521 | 0.389 | 0.403 | ... | 0.176 | 0.284 | 0.197 | 1 | 0.346 |
| 1638 | 0.409 | 0.179 | 0.638 | 0.327 | 0.405 | ... | 0.182 | 0.193 | 0.132 | 0.346 | Inf |




(b) Jiang and Conrath Semantic Similarity

| Entrez ID | 2306 | 9447 | 6656 | 4249 | 6187 | ... | 8658 | 7157 | 2747 | 5118 | 1638 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2306 | 1 | 0.676 | 0.815 | 0.128 | 0.44 | ... | 0.389 | 0.659 | 0.196 | 0.329 | 0.176 |
| 9447 | 0.676 | 1 | 0.723 | 0.128 | 0.678 | ... | 0.598 | 0.664 | 0.225 | 0.505 | 0.237 |
| 6656 | 0.815 | 0.723 | 1 | 0.128 | 0.541 | ... | 0.473 | 0.735 | 0.224 | 0.401 | 0.207 |
| 4249 | 0.128 | 0.128 | 0.128 | 1 | 0.128 | ... | 0.411 | 0.128 | 0.16 | 0.103 | 0.16 |
| 6187 | 0.44 | 0.678 | 0.541 | 0.128 | 1 | ... | 0.571 | 0.586 | 0.203 | 0.483 | 0.232 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 8658 | 0.389 | 0.598 | 0.473 | 0.411 | 0.571 | ... | 1 | 0.635 | 0.258 | 0.423 | 0.442 |
| 7157 | 0.659 | 0.664 | 0.735 | 0.128 | 0.586 | ... | 0.635 | 1 | 0.303 | 0.471 | 0.287 |
| 2747 | 0.196 | 0.223 | 0.224 | 0.16 | 0.203 | ... | 0.258 | 0.303 | 1 | 0.25 | 0.25 |
| 5118 | 0.329 | 0.505 | 0.401 | 0.103 | 0.483 | ... | 0.423 | 0.471 | 0.25 | 1 | 0.212 |
| 1638 | 0.176 | 0.237 | 0.207 | 0.16 | 0.232 | ... | 0.442 | 0.287 | 0.25 | 0.212 | 1 |

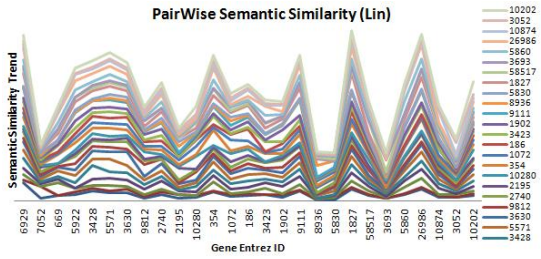



(c) Wang Semantic Similarity

**Figure A-3:** Pair-Wise Semantic Similarity Matrix using (a) Lin, (b) Jiang and Conrath and (c) Wang for Lymphoma dataset

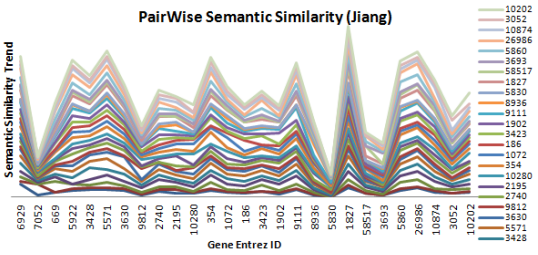
*Figure* A-4 below shows the pair-wise semantic similarity values and plots of some of the genes for the Lin, Jiang and Conrath and Wang measures for the Embryonal Tumours of the Central Nervous System dataset.

| Entrez ID | 6929 | 7052 | 669 | 5922 | 3428 | ... | 678 | 1609 | 5108 | 3796 | 273 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6929 | 1.461 | 0.224 | 0.469 | 0.53 | 0.856 | ... | 0.982 | 0.286 | 0.396 | 0.316 | 0.132 |
| 7052 | 0.224 | 1 | 0 | 0.185 | 0.163 | ... | 0.088 | 0.141 | 0.172 | 0.078 | 0.423 |
| 669 | 0.469 | 0 | 1 | 0.296 | 0.263 | ... | 0.36 | 0.085 | 0.064 | 0.075 | 0.134 |
| 5922 | 0.53 | 0.185 | 0.296 | Inf | 0.49 | ... | 0.423 | 0.423 | 0.176 | 0.193 | 0.323 |
| 3428 | 0.856 | 0.163 | 0.263 | 0.49 | 1.164 | ... | 0.725 | 0.326 | 0.26 | 0.283 | 0.133 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 678 | 0.982 | 0.088 | 0.36 | 0.423 | 0.725 | ... | 1.083 | 0.193 | 0.303 | 0.182 | 0.136 |
| 1609 | 0.286 | 0.141 | 0.085 | 0.423 | 0.326 | ... | 0.193 | 1.026 | 0.198 | 0.359 | 0.276 |
| 5108 | 0.396 | 0.172 | 0.064 | 0.176 | 0.26 | ... | 0.303 | 0.198 | 1 | 0.552 | 0.221 |
| 3796 | 0.316 | 0.078 | 0.075 | 0.193 | 0.283 | ... | 0.182 | 0.359 | 0.552 | 1 | 0.133 |
| 273 | 0.132 | 0.423 | 0.134 | 0.323 | 0.133 | ... | 0.136 | 0.276 | 0.221 | 0.133 | 1 |



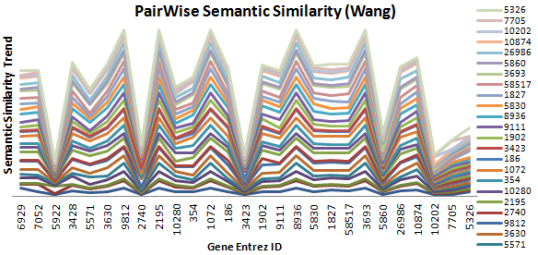

(a) Lin Semantic Similarity

| Entrez ID | 6929 | 7052 | 669 | 5922 | 3428 | ... | 678 | 1609 | 5108 | 3796 | 273 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6929 | 1.255 | 0.189 | 0.421 | 0.594 | 0.73 | ... | 0.809 | 0.321 | 0.362 | 0.318 | 0.196 |
| 7052 | 0.189 | 1 | 0.024 | 0.202 | 0.1 | ... | 0.077 | 0.081 | 0.034 | 0.073 | 0.408 |
| 669 | 0.421 | 0.024 | 1 | 0.417 | 0.261 | ... | 0.336 | 0.11 | 0.069 | 0.152 | 0.21 |
| 5922 | 0.594 | 0.202 | 0.417 | Inf | 0.562 | ... | 0.507 | 0.491 | 0.271 | 0.331 | 0.443 |
| 3428 | 0.73 | 0.1 | 0.261 | 0.562 | 1.105 | ... | 0.66 | 0.326 | 0.236 | 0.302 | 0.189 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 678 | 0.809 | 0.077 | 0.336 | 0.507 | 0.66 | ... | 1.007 | 0.234 | 0.307 | 0.241 | 0.214 |
| 1609 | 0.321 | 0.081 | 0.11 | 0.491 | 0.326 | ... | 0.234 | 1.039 | 0.158 | 0.342 | 0.297 |
| 5108 | 0.362 | 0.034 | 0.069 | 0.271 | 0.236 | ... | 0.307 | 0.158 | 1 | 0.499 | 0.179 |
| 3796 | 0.318 | 0.073 | 0.152 | 0.331 | 0.302 | ... | 0.241 | 0.342 | 0.499 | 1 | 0.224 |
| 273 | 0.196 | 0.408 | 0.21 | 0.443 | 0.189 | ... | 0.214 | 0.297 | 0.179 | 0.224 | 1 |




(b) Jiang and Conrath Semantic Similarity

| Entrez ID | 6929 | 7052 | 5922 | 3428 | 5571 | ... | 678 | 1609 | 5108 | 3796 | 273 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6929 | 1 | 0.507 | 0.135 | 0.69 | 0.455 | ... | 0.438 | 0.56 | 0.557 | 0.505 | 1 |
| 7052 | 0.507 | 1 | 0.135 | 0.654 | 0.396 | ... | 0.269 | 0.669 | 0.743 | 0.578 | 1 |
| 5922 | 0.135 | 0.135 | 1 | 0.135 | 0.212 | ... | 0.135 | 0.135 | 0.135 | 0.135 | 0.135 |
| 3428 | 0.69 | 0.654 | 0.135 | 1 | 0.405 | ... | 0.373 | 0.688 | 0.786 | 0.648 | 1 |
| 5571 | 0.455 | 0.396 | 0.212 | 0.405 | 1 | ... | 0.211 | 0.609 | 0.406 | 0.344 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 678 | 0.438 | 0.269 | 0.135 | 0.373 | 0.211 | ... | 1 | 0.245 | 0.296 | 0.262 | 0.328 |
| 1609 | 0.56 | 0.669 | 0.135 | 0.688 | 0.609 | ... | 0.245 | 1 | 0.718 | 0.621 | 1 |
| 5108 | 0.557 | 0.743 | 0.135 | 0.786 | 0.406 | ... | 0.296 | 0.718 | 1 | 0.736 | 1 |
| 3796 | 0.505 | 0.578 | 0.135 | 0.648 | 0.344 | ... | 0.262 | 0.621 | 0.736 | 1 | 1 |
| 273 | 1 | 1 | 0.135 | 1 | 1 | ... | 0.328 | 1 | 1 | 1 | 1 |



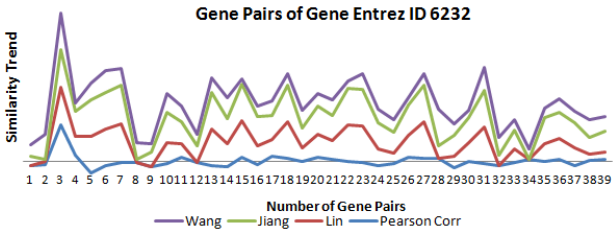

(c) Wang Semantic Similarity

**Figure A-4:** Pair-Wise Semantic Similarity Matrix using (a) Lin, (b) Jiang and Conrath and (c) Wang for Embryonal Tumours of the Central Nervous System

## C. Comparison of Pair-wise Gene Expression Similarity and Semantic Similarity

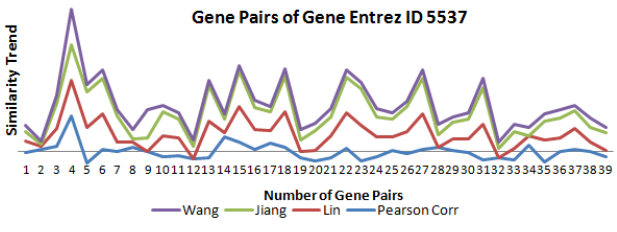
The *Figure* A-5 gives a comparison of expression similarity and semantic similarity of four sample gene pairs of the breast cancer dataset, suggesting that gene products with similar expression patterns may have similarly annotated profiles.

| ENTREZ ID | 6232 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| 54704 | -0.1226 | 0.000 | 0.267 | 0.301 |
| 11014 | -0.0989 | 0.089 | 0.045 | 0.691 |
| 6232 | 1.0000 | 1.024 | 1.012 | 1.000 |
| 5547 | 0.1653 | 0.498 | 0.692 | 0.225 |
| 23641 | -0.3178 | 1.000 | 1.000 | 0.436 |
| ... | ... | ... | ... | ... |
| 8573 | -0.0128 | 0.498 | 0.707 | 0.260 |
| 1112 | 0.0442 | 0.558 | 0.722 | 0.374 |
| 1644 | -0.1371 | 0.498 | 0.684 | 0.327 |
| 29089 | 0.0138 | 0.183 | 0.457 | 0.484 |
| 10112 | 0.0367 | 0.213 | 0.559 | 0.399 |



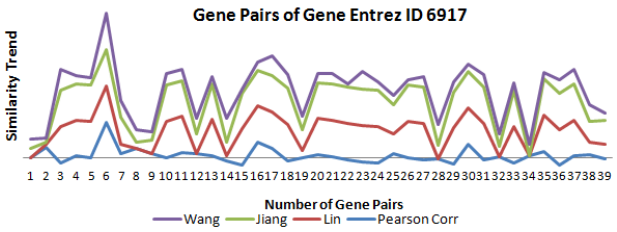

(a) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 6232

| ENTREZ ID | 5547 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| 54704 | -0.0127 | 0.320 | 0.267 | 0.182 |
| 11014 | 0.0827 | 0.089 | 0.045 | 0.097 |
| 6232 | 0.1653 | 0.498 | 0.692 | 0.225 |
| 5547 | 1.0000 | 1.000 | 1.000 | 1.000 |
| 23641 | -0.3031 | 1.000 | 1.000 | 0.182 |
| ... | ... | ... | ... | ... |
| 8573 | -0.2679 | 0.594 | 0.537 | 0.215 |
| 1112 | 0.0165 | 0.382 | 0.559 | 0.223 |
| 1644 | 0.0637 | 0.585 | 0.509 | 0.130 |
| 29089 | -0.0001 | 0.262 | 0.412 | 0.277 |
| 10112 | -0.1489 | 0.201 | 0.481 | 0.143 |



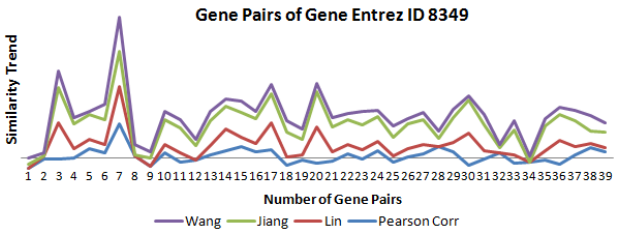

(b) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 5547

| ENTREZ ID | 6917 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **54704** | 0.0061 | 0.000 | 0.267 | 0.268 |
| **11014** | 0.2997 | 0.089 | 0.045 | 0.138 |
| **6232** | -0.1181 | 1.000 | 1.000 | 0.588 |
| **5547** | 0.0587 | 1.000 | 1.000 | 0.232 |
| **23641** | 0.0262 | 1.000 | 1.000 | 0.203 |
| **…** | … | … | … | … |
| **8573** | 0.1852 | 1.000 | 1.000 | 0.186 |
| **1112** | -0.1999 | 1.000 | 1.000 | 0.368 |
| **1644** | 0.0581 | 1.000 | 1.000 | 0.394 |
| **29089** | 0.0855 | 0.346 | 0.574 | 0.482 |
| **10112** | -0.0124 | 0.402 | 0.665 | 0.195 |




(c) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 6917

| ENTREZ ID | 8349 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **54704** | -0.3125 | 0.000 | 0.129 | 0.176 |
| **11014** | -0.0272 | 0.079 | 0.000 | 0.098 |
| **6232** | -0.0473 | 1.071 | 1.035 | 0.485 |
| **5547** | 0.0101 | 0.257 | 0.730 | 0.182 |
| **23641** | 0.2711 | 0.257 | 0.730 | 0.114 |
| **…** | … | … | … | … |
| **8573** | -0.0644 | 0.257 | 0.730 | 0.216 |
| **1112** | -0.1827 | 0.693 | 0.763 | 0.198 |
| **1644** | 0.0868 | 0.257 | 0.730 | 0.315 |
| **29089** | 0.3001 | 0.127 | 0.359 | 0.456 |
| **10112** | 0.1703 | 0.134 | 0.456 | 0.260 |

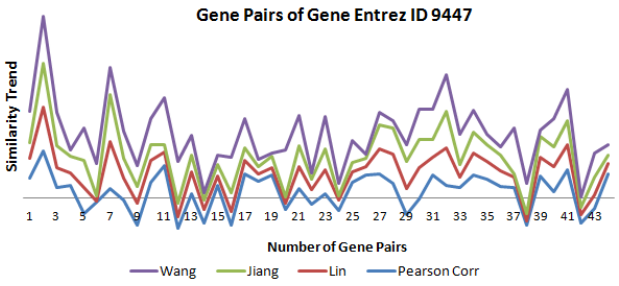



(d) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 8349

**Figure A-5:** Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of sample gene pairs of Breast Cancer dataset

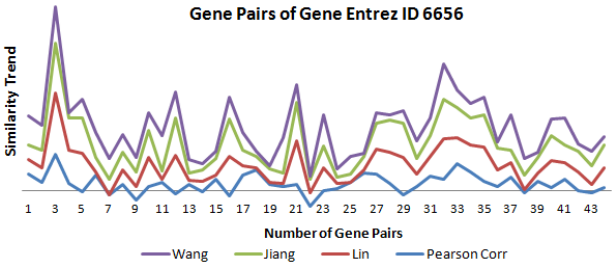
*Figure* A-6 and *Figure* A-7 show the comparison of expression similarity and semantic similarity of four sample gene pairs of lymphoma dataset and embryonal tumours of central nervous system dataset. The assumption that gene products with similar expression patterns may have similarly annotated profiles also seems to hold true for this dataset. The graph obtained from the plot of the values for the genes exhibit a similar trend.

| ENTREZ ID | 9447 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **2306** | 0.423325 | 0.428 | 0.325 | 0.676 |
| **9447** | 1 | 0.944 | 0.946 | 1 |
| **6656** | 0.213661 | 0.423 | 0.471 | 0.723 |
| **4249** | 0.273306 | 0.263 | 0.351 | 0.128 |
| **6187** | -0.33553 | 0.555 | 0.589 | 0.678 |
| **…** | … | … | … | … |
| **8658** | 0.130979 | 0.544 | 0.412 | 0.598 |
| **7157** | 0.602441 | 0.541 | 0.514 | 0.664 |
| **2747** | -0.53223 | 0.174 | 0.156 | 0.223 |
| **5118** | -0.23585 | 0.301 | 0.381 | 0.505 |
| **1638** | 0.507952 | 0.218 | 0.179 | 0.237 |



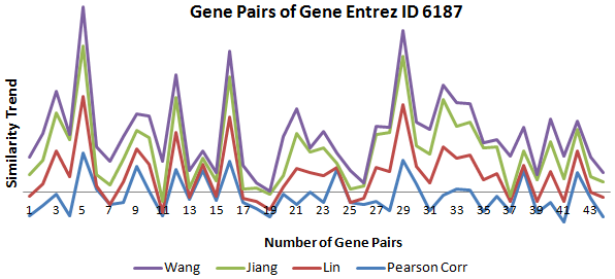

(a) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 9447

| ENTREZ ID | 6656 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **2306** | 0.447446 | 0.407 | 0.412 | 0.815 |
| **9447** | 0.213661 | 0.423 | 0.471 | 0.723 |
| **6656** | 1 | 1.73 | 1.387 | 1 |
| **4249** | 0.186839 | 0.923 | 0.914 | 0.128 |
| **6187** | -0.03163 | 1.07 | 0.977 | 0.541 |
| … | … | … | … | … |
| **8658** | 0.062237 | 0.756 | 0.707 | 0.473 |
| **7157** | 0.292448 | 0.481 | 0.504 | 0.735 |
| **2747** | -0.02104 | 0.527 | 0.578 | 0.224 |
| **5118** | -0.06807 | 0.236 | 0.521 | 0.401 |
| **1638** | 0.083528 | 0.555 | 0.638 | 0.207 |



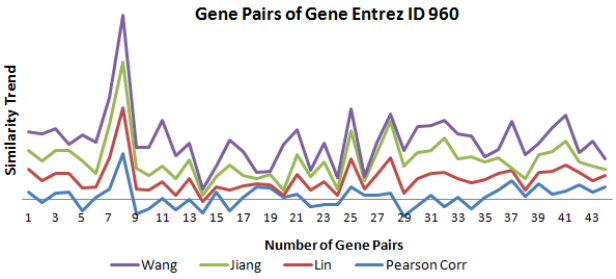

(b) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 6656

| ENTREZ ID | 6187 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **2306** | -0.58725 | 0.504 | 0.533 | 0.44 |
| **9447** | -0.33553 | 0.555 | 0.589 | 0.678 |
| **6656** | -0.03163 | 1.07 | 0.977 | 0.541 |
| **4249** | -0.58127 | 0.979 | 0.931 | 0.128 |
| **6187** | 1 | 1.432 | 1.254 | 1 |
| … | … | … | … | … |
| **8658** | -0.24991 | 0.783 | 0.761 | 0.571 |
| **7157** | -0.7338 | 0.513 | 0.56 | 0.586 |
| **2747** | 0.501957 | 0.549 | 0.541 | 0.203 |
| **5118** | -0.14327 | 0.151 | 0.403 | 0.483 |
| **1638** | -0.61011 | 0.485 | 0.405 | 0.232 |




(c) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 6187

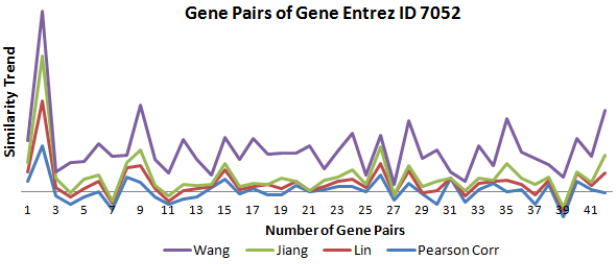
| ENTREZ ID | 960 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **2306** | 0.174881 | 0.486 | 0.405 | 0.401 |
| **9447** | -0.05891 | 0.485 | 0.419 | 0.581 |
| **6656** | 0.151369 | 0.425 | 0.485 | 0.486 |
| **4249** | 0.155759 | 0.412 | 0.501 | 0.138 |
| **6187** | -0.24843 | 0.51 | 0.58 | 0.561 |
| … | … | … | … | … |
| **8658** | 0.131491 | 0.495 | 0.432 | 0.508 |
| **7157** | 0.193857 | 0.557 | 0.525 | 0.566 |
| **2747** | 0.313379 | 0.274 | 0.236 | 0.203 |
| **5118** | 0.171897 | 0.24 | 0.321 | 0.542 |
| **1638** | 0.276203 | 0.254 | 0.133 | 0.221 |




(d) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 960

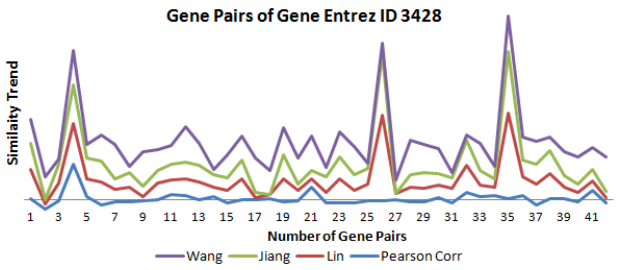
**Figure A-6:** Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of sample gene pairs of Lymphoma dataset

| ENTREZ ID | 7052 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **6929** | 0.217468 | 0.224 | 0.189 | 0.507 |
| **7052** | 1 | 1 | 1 | 1 |
| **5922** | -0.09129 | 0.185 | 0.202 | 0.135 |
| **3428** | -0.28663 | 0.163 | 0.1 | 0.654 |
| **5571** | -0.12827 | 0.198 | 0.203 | 0.396 |
| … | … | … | … | … |
| **678** | 0.160242 | 0.088 | 0.077 | 0.269 |
| **1609** | -0.56609 | 0.141 | 0.081 | 0.669 |
| **5108** | 0.227934 | 0.172 | 0.034 | 0.743 |
| **3796** | 0.045173 | 0.078 | 0.073 | 0.578 |
| **273** | -0.02322 | 0.423 | 0.408 | 1 |



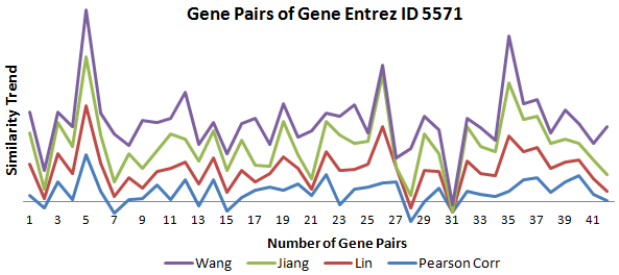

(a) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 7052

| ENTREZ ID | 3428 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **6929** | 0.000704 | 0.856 | 0.73 | 0.69 |
| **7052** | -0.28663 | 0.163 | 0.1 | 0.654 |
| **5922** | -0.03786 | 0.49 | 0.562 | 0.135 |
| **3428** | 1 | 1.164 | 1.105 | 1 |
| **5571** | 0.063185 | 0.533 | 0.583 | 0.405 |
| … | … | … | … | … |
| **678** | 0.009639 | 0.725 | 0.66 | 0.373 |
| **1609** | 0.026001 | 0.326 | 0.326 | 0.688 |
| **5108** | -0.06424 | 0.26 | 0.236 | 0.786 |
| **3796** | 0.25553 | 0.283 | 0.302 | 0.648 |
| **273** | -0.10085 | 0.133 | 0.189 | 1 |



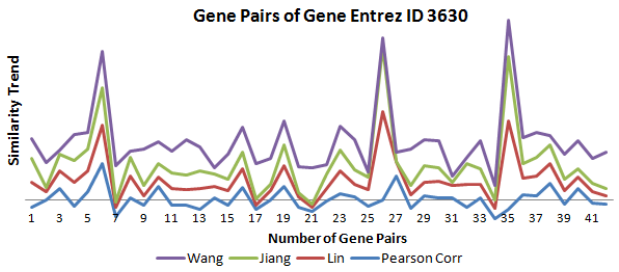

(b) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 3428

| ENTREZ ID | 5571 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **6929** | 0.152185 | 0.645 | 0.659 | 0.455 |
| **7052** | -0.12827 | 0.198 | 0.203 | 0.396 |
| **5922** | 0.431054 | 0.59 | 0.669 | 0.212 |
| **3428** | 0.063185 | 0.533 | 0.583 | 0.405 |
| **5571** | 1 | 1.03 | 1.029 | 1 |
| … | … | … | … | … |
| **678** | 0.212264 | 0.496 | 0.542 | 0.211 |
| **1609** | 0.435734 | 0.402 | 0.492 | 0.609 |
| **5108** | 0.554288 | 0.344 | 0.35 | 0.406 |
| **3796** | 0.154019 | 0.33 | 0.414 | 0.344 |
| **273** | 0.035894 | 0.204 | 0.348 | 1 |




(c) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 5571

| ENTREZ ID | 3630 | | | |
|---|---|---|---|---|
| | Pearson Corr | Semantic | | |
| | | Lin | Jiang | Wang |
| **6929** | -0.18513 | 0.692 | 0.641 | 0.559 |
| **7052** | -0.00921 | 0.243 | 0.123 | 0.69 |
| **5922** | 0.324569 | 0.473 | 0.467 | 0.135 |
| **3428** | -0.17857 | 0.668 | 0.613 | 0.724 |
| **5571** | 0.227748 | 0.596 | 0.595 | 0.452 |
| … | … | … | … | … |
| **678** | 0.478898 | 0.531 | 0.516 | 0.253 |
| **1609** | -0.11685 | 0.373 | 0.332 | 0.692 |
| **5108** | 0.330671 | 0.317 | 0.222 | 0.784 |
| **3796** | -0.07251 | 0.295 | 0.256 | 0.686 |
| **273** | -0.09783 | 0.211 | 0.217 | 1 |



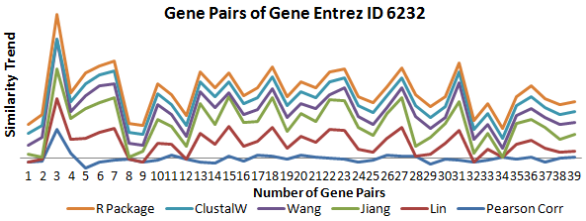

(d) Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of Gene 3630

**Figure A-7:** Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang of sample gene pairs of Embryonal Tumours of Central Nervous System dataset

## D. Comparison of Pair-wise Gene Expression Similarity, Semantic Similarity and Sequence Similarity

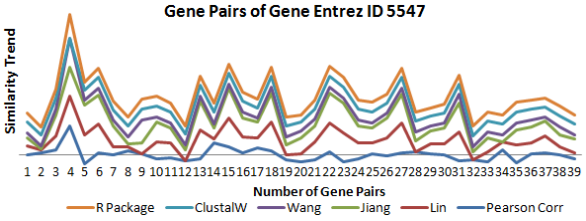
The gene expression similarity, Lin, Jiang and Conrath and Wang semantic similarity and sequence similarity of the some of the gene pairs of breast cancer dataset is given in *Figure* A-8. From the graph it can be clearly observed that the genes pair-wise scores for the various measures follow a common trend, indicating a correlation between gene expression similarity, semantic similarity and sequence similarity.

| ENTREZ ID | 6232 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| 54704 | -0.1226 | 0.000 | 0.267 | 0.301 | 0.3964 | 0.3255 |
| 11014 | -0.0989 | 0.089 | 0.045 | 0.691 | 0.4 | 0.3533 |
| 6232 | 1.0000 | 1.024 | 1.012 | 1.000 | | 0.8471 |
| 5547 | 0.1653 | 0.498 | 0.692 | 0.225 | 0.3548 | 0.3073 |
| 23641 | -0.3178 | 1.000 | 1.000 | 0.436 | 0.4012 | 0.3672 |
| … | … | … | … | … | … | … |
| 8573 | -0.0128 | 0.498 | 0.707 | 0.260 | 0.3405 | 0.3134 |
| 1112 | 0.0442 | 0.558 | 0.722 | 0.374 | 0.3964 | 0.3574 |
| 1644 | -0.1371 | 0.498 | 0.684 | 0.327 | 0.3405 | 0.301 |
| 29089 | 0.0138 | 0.183 | 0.457 | 0.484 | 0.3619 | 0.3227 |
| 10112 | 0.0367 | 0.213 | 0.559 | 0.399 | 0.3935 | 0.3135 |



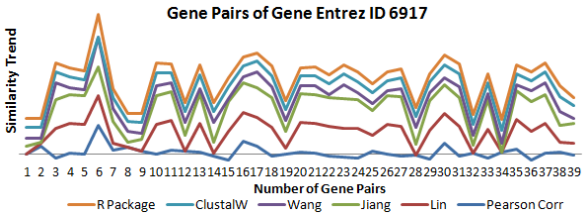

(a) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 6232

| ENTREZ ID | 5547 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| 54704 | -0.0127 | 0.320 | 0.267 | 0.182 | 0.3738 | 0.3087 |
| 11014 | 0.0827 | 0.089 | 0.045 | 0.097 | 0.3536 | 0.2977 |
| 6232 | 0.1653 | 0.498 | 0.692 | 0.225 | 0.3548 | 0.3073 |
| 5547 | 1.0000 | 1.000 | 1.000 | 1.000 | | 0.803 |
| 23641 | -0.3031 | 1.000 | 1.000 | 0.182 | 0.344 | 0.279 |
| … | … | … | … | … | … | … |
| 8573 | -0.2679 | 0.594 | 0.537 | 0.215 | 0.3857 | 0.328 |
| 1112 | 0.0165 | 0.382 | 0.559 | 0.223 | 0.3869 | 0.3121 |
| 1644 | 0.0637 | 0.585 | 0.509 | 0.130 | 0.3524 | 0.308 |
| 29089 | -0.0001 | 0.262 | 0.412 | 0.277 | 0.3643 | 0.3396 |
| 10112 | -0.1489 | 0.201 | 0.481 | 0.143 | 0.3935 | 0.2941 |



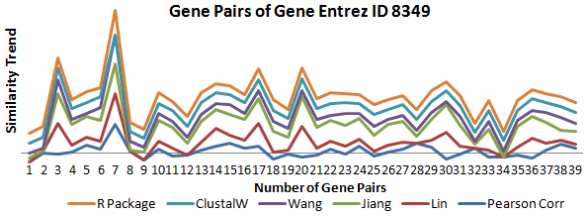

(b) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 5547

| ENTREZ ID | 6917 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| 54704 | 0.0061 | 0.000 | 0.267 | 0.268 | 0.3786 | 0.3305 |
| 11014 | 0.2997 | 0.089 | 0.045 | 0.138 | 0.3583 | 0.3096 |
| 6232 | -0.1181 | 1.000 | 1.000 | 0.588 | 0.3643 | 0.3 |
| 5547 | 0.0587 | 1.000 | 1.000 | 0.232 | 0.3714 | 0.3163 |
| 23641 | 0.0262 | 1.000 | 1.000 | 0.203 | 0.3333 | 0.3023 |
| … | … | … | … | … | … | … |
| 8573 | 0.1852 | 1.000 | 1.000 | 0.186 | 0.3738 | 0.3276 |
| 1112 | -0.1999 | 1.000 | 1.000 | 0.368 | 0.3583 | 0.2987 |
| 1644 | 0.0581 | 1.000 | 1.000 | 0.394 | 0.3845 | 0.3094 |
| 29089 | 0.0855 | 0.346 | 0.574 | 0.482 | 0.4964 | 0.3806 |
| 10112 | -0.0124 | 0.402 | 0.665 | 0.195 | 0.4182 | 0.2905 |




(c) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 6917

| ENTREZ ID | 8349 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| 54704 | -0.3125 | 0.000 | 0.129 | 0.176 | 0.3524 | 0.3422 |
| 11014 | -0.0272 | 0.079 | 0.000 | 0.098 | 0.4 | 0.3612 |
| 6232 | -0.0473 | 1.071 | 1.035 | 0.485 | 0.4119 | 0.3563 |
| 5547 | 0.0101 | 0.257 | 0.730 | 0.182 | 0.3595 | 0.3111 |
| 23641 | 0.2711 | 0.257 | 0.730 | 0.114 | 0.3952 | 0.3581 |
| … | … | … | … | … | … | … |
| 8573 | -0.0644 | 0.257 | 0.730 | 0.216 | 0.3524 | 0.3226 |
| 1112 | -0.1827 | 0.693 | 0.763 | 0.198 | 0.4024 | 0.3387 |
| 1644 | 0.0868 | 0.257 | 0.730 | 0.315 | 0.3679 | 0.301 |
| 29089 | 0.3001 | 0.127 | 0.359 | 0.456 | 0.3833 | 0.3302 |
| 10112 | 0.1703 | 0.134 | 0.456 | 0.260 | 0.4 | 0.3173 |



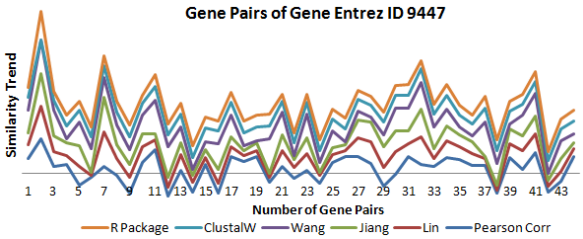

(d) Comparison of Gene Expression Similarity, Semantic Similarity Sequence Similarity for Lin, Jiang and Conrath and Wang and of Gene 8349

**Figure A-8:** Comparison of Gene Expression Similarity, Semantic Similarity and Sequence Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of sample gene pairs of Breast Cancer dataset

*Figure* A-9 depicts the gene expression similarity, Lin, Jiang and Conrath and Wang semantic similarity and sequence similarity of the some of the gene pairs of Lymphoma dataset. From the graph it can be clearly observed that the genes pair-wise scores for the various measures follow a common trend, indicating a correlation between gene expression similarity, semantic similarity and sequence similarity. This

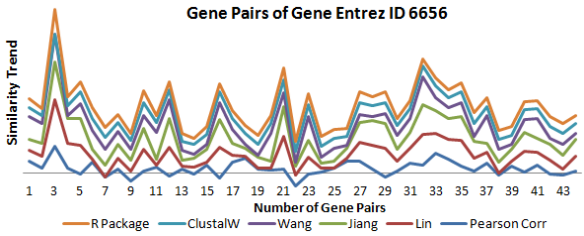
is borne out also by *Figure* A-10 for the gene pairs of Embryonal Tumours of Central Nervous System dataset.

| ENTREZ ID | 9447 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| 2306 | 0.423325 | 0.428 | 0.325 | 0.676 | 0.3631 | 0.2828 |
| 9447 | 1 | 0.944 | 0.946 | 1 | | 0.8082 |
| 6656 | 0.213661 | 0.423 | 0.471 | 0.723 | 0.306 | 0.2394 |
| 4249 | 0.273306 | 0.263 | 0.351 | 0.128 | 0.3679 | 0.3092 |
| 6187 | -0.33553 | 0.555 | 0.589 | 0.678 | 0.3429 | 0.2649 |
| ... | ... | ... | ... | ... | ... | ... |
| 8658 | 0.130979 | 0.544 | 0.412 | 0.598 | 0.3417 | 0.2854 |
| 7157 | 0.602441 | 0.541 | 0.514 | 0.664 | 0.3452 | 0.3049 |
| 2747 | -0.53223 | 0.174 | 0.156 | 0.223 | 0.3488 | 0.2572 |
| 5118 | -0.23585 | 0.301 | 0.381 | 0.505 | 0.3571 | 0.2796 |
| 1638 | 0.507952 | 0.218 | 0.179 | 0.237 | 0.3738 | 0.3146 |



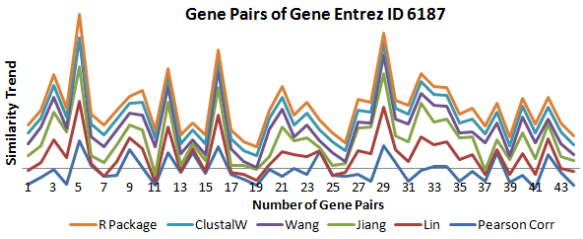

(a) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 9447

| ENTREZ ID | 6656 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| 2306 | 0.447446 | 0.407 | 0.412 | 0.815 | 0.3619 | 0.309 |
| 9447 | 0.213661 | 0.423 | 0.471 | 0.723 | 0.306 | 0.2394 |
| 6656 | 1 | 1.73 | 1.387 | 1 | | 0.9203 |
| 4249 | 0.186839 | 0.923 | 0.914 | 0.128 | 0.3595 | 0.3037 |
| 6187 | -0.03163 | 1.07 | 0.977 | 0.541 | 0.4452 | 0.3875 |
| ... | ... | ... | ... | ... | ... | ... |
| 8658 | 0.062237 | 0.756 | 0.707 | 0.473 | 0.3536 | 0.284 |
| 7157 | 0.292448 | 0.481 | 0.504 | 0.735 | 0.3607 | 0.3285 |
| 2747 | -0.02104 | 0.527 | 0.578 | 0.224 | 0.4286 | 0.364 |
| 5118 | -0.06807 | 0.236 | 0.521 | 0.401 | 0.3952 | 0.3457 |
| 1638 | 0.083528 | 0.555 | 0.638 | 0.207 | 0.356 | 0.2987 |



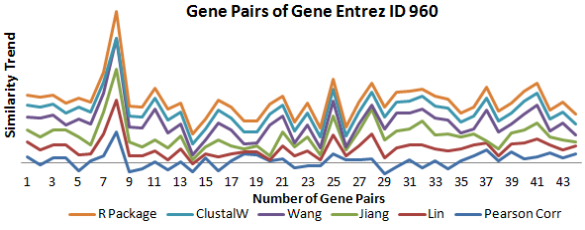

(b) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 6656

| ENTREZ ID | 6187 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| 2306 | -0.58725 | 0.504 | 0.533 | 0.44 | 0.375 | 0.3097 |
| 9447 | -0.33553 | 0.555 | 0.589 | 0.678 | 0.3429 | 0.2649 |
| 6656 | -0.03163 | 1.07 | 0.977 | 0.541 | 0.4452 | 0.3875 |
| 4249 | -0.58127 | 0.979 | 0.931 | 0.128 | 0.3869 | 0.3269 |
| 6187 | 1 | 1.432 | 1.254 | 1 | | 0.8659 |
| ... | ... | ... | ... | ... | ... | ... |
| 8658 | -0.24991 | 0.783 | 0.761 | 0.571 | 0.3738 | 0.3059 |
| 7157 | -0.7338 | 0.513 | 0.56 | 0.586 | 0.3548 | 0.3201 |
| 2747 | 0.501957 | 0.549 | 0.541 | 0.203 | 0.425 | 0.3614 |
| 5118 | -0.14327 | 0.151 | 0.403 | 0.483 | 0.419 | 0.3374 |
| 1638 | -0.61011 | 0.485 | 0.405 | 0.232 | 0.3524 | 0.3052 |




(c) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 6187

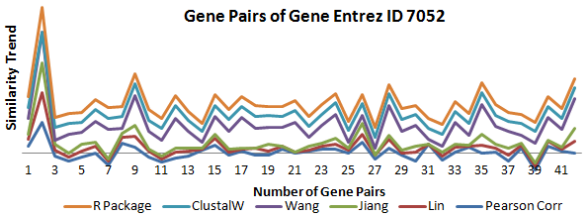
| ENTREZ ID | 960 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| 2306 | 0.174881 | 0.486 | 0.405 | 0.401 | 0.381 | 0.316 |
| 9447 | -0.05891 | 0.485 | 0.419 | 0.581 | 0.3714 | 0.3081 |
| 6656 | 0.151369 | 0.425 | 0.485 | 0.486 | 0.344 | 0.2876 |
| 4249 | 0.155759 | 0.412 | 0.501 | 0.138 | 0.3869 | 0.3383 |
| 6187 | -0.24843 | 0.51 | 0.58 | 0.561 | 0.3774 | 0.3084 |
| ... | ... | ... | ... | ... | ... | ... |
| 8658 | 0.131491 | 0.495 | 0.432 | 0.508 | 0.4143 | 0.3317 |
| 7157 | 0.193857 | 0.557 | 0.525 | 0.566 | 0.3976 | 0.3313 |
| 2747 | 0.313379 | 0.274 | 0.236 | 0.203 | 0.3667 | 0.3059 |
| 5118 | 0.171897 | 0.24 | 0.321 | 0.542 | 0.3548 | 0.3183 |
| 1638 | 0.276203 | 0.254 | 0.133 | 0.221 | 0.3667 | 0.3245 |




(d) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 960

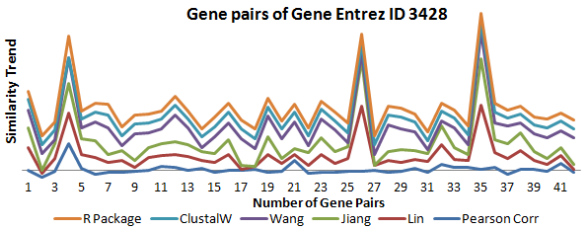
**Figure A-9:** Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of sample gene pairs of Lymphoma dataset

| ENTREZ ID | 7052 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| **6929** | 0.217468 | 0.224 | 0.189 | 0.507 | 0.3786 | 0.3518 |
| **7052** | 1 | 1 | 1 | 1 | | 0.8407 |
| **5922** | -0.09129 | 0.185 | 0.202 | 0.135 | 0.4119 | 0.3407 |
| **3428** | -0.28663 | 0.163 | 0.1 | 0.654 | 0.3476 | 0.3301 |
| **5571** | -0.12827 | 0.198 | 0.203 | 0.396 | 0.3583 | 0.3028 |
| … | … | … | … | … | … | … |
| **678** | 0.160242 | 0.088 | 0.077 | 0.269 | 0.3786 | 0.316 |
| **1609** | -0.56609 | 0.141 | 0.081 | 0.669 | 0.3821 | 0.3107 |
| **5108** | 0.227934 | 0.172 | 0.034 | 0.743 | 0.3667 | 0.3354 |
| **3796** | 0.045173 | 0.078 | 0.073 | 0.578 | 0.3571 | 0.3342 |
| **273** | -0.02322 | 0.423 | 0.408 | 1 | 0.35 | 0.3119 |



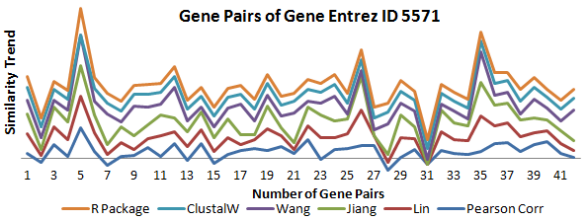

(a) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 7052

| ENTREZ ID | 3428 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| **6929** | 0.000704 | 0.856 | 0.73 | 0.69 | 0.3929 | 0.3127 |
| **7052** | -0.28663 | 0.163 | 0.1 | 0.654 | 0.3476 | 0.3301 |
| **5922** | -0.03786 | 0.49 | 0.562 | 0.135 | 0.356 | 0.3208 |
| **3428** | 1 | 1.164 | 1.105 | 1 | | 0.8271 |
| **5571** | 0.063185 | 0.533 | 0.583 | 0.405 | 0.3667 | 0.2846 |
| … | … | … | … | … | … | … |
| **678** | 0.009639 | 0.725 | 0.66 | 0.373 | 0.3619 | 0.2988 |
| **1609** | 0.026001 | 0.326 | 0.326 | 0.688 | 0.3607 | 0.2943 |
| **5108** | -0.06424 | 0.26 | 0.236 | 0.786 | 0.3714 | 0.3311 |
| **3796** | 0.25553 | 0.283 | 0.302 | 0.648 | 0.3631 | 0.3124 |
| **273** | -0.10085 | 0.133 | 0.189 | 1 | 0.35 | 0.3182 |



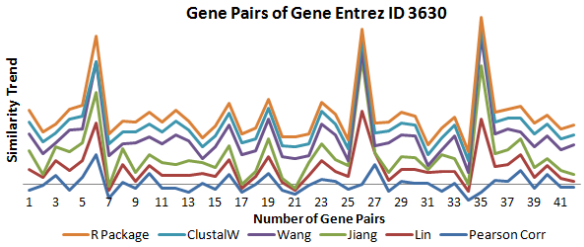

(b) Comparison of Gene Expression Similarity, Semantic Similarity and Sequence Similarity for Lin, Jiang and Conrath and Wang of Gene 3428

| ENTREZ ID | 5571 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| **6929** | 0.152185 | 0.645 | 0.659 | 0.455 | 0.4155 | 0.3561 |
| **7052** | -0.12827 | 0.198 | 0.203 | 0.396 | 0.3583 | 0.3028 |
| **5922** | 0.431054 | 0.59 | 0.669 | 0.212 | 0.3619 | 0.2628 |
| **3428** | 0.063185 | 0.533 | 0.583 | 0.405 | 0.3667 | 0.2846 |
| **5571** | 1 | 1.03 | 1.029 | 1 | | 0.8731 |
| … | … | … | … | … | … | … |
| **678** | 0.212264 | 0.496 | 0.542 | 0.211 | 0.4298 | 0.3711 |
| **1609** | 0.435734 | 0.402 | 0.492 | 0.609 | 0.4036 | 0.3201 |
| **5108** | 0.554288 | 0.344 | 0.35 | 0.406 | 0.35 | 0.2806 |
| **3796** | 0.154019 | 0.33 | 0.414 | 0.344 | 0.3726 | 0.3031 |
| **273** | 0.035894 | 0.204 | 0.348 | 1 | 0.3798 | 0.3067 |




(c) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 5571

| ENTREZ ID | 3630 | | | | | |
|---|---|---|---|---|---|---|
| | Pearson Corr | Semantic | | | Sequence | |
| | | Lin | Jiang | Wang | ClustalW | R Package |
| **6929** | -0.18513 | 0.692 | 0.641 | 0.559 | 0.4107 | 0.3726 |
| **7052** | -0.00921 | 0.243 | 0.123 | 0.69 | 0.3893 | 0.3297 |
| **5922** | 0.324569 | 0.473 | 0.467 | 0.135 | 0.3512 | 0.2692 |
| **3428** | -0.17857 | 0.668 | 0.613 | 0.724 | 0.3774 | 0.3271 |
| **5571** | 0.227748 | 0.596 | 0.595 | 0.452 | 0.4333 | 0.3697 |
| … | … | … | … | … | … | … |
| **678** | 0.478898 | 0.531 | 0.516 | 0.253 | 0.4464 | 0.3986 |
| **1609** | -0.11685 | 0.373 | 0.332 | 0.692 | 0.4214 | 0.355 |
| **5108** | 0.330671 | 0.317 | 0.222 | 0.784 | 0.375 | 0.3003 |
| **3796** | -0.07251 | 0.295 | 0.256 | 0.686 | 0.3821 | 0.328 |
| **273** | -0.09783 | 0.211 | 0.217 | 1 | 0.35 | 0.3156 |




(d) Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of Gene 3630

**Figure A-10:** Comparison of Gene Expression Similarity, Semantic Similarity for Lin, Jiang and Conrath and Wang and Sequence Similarity of sample gene pairs of Embryonal Tumours of Central Nervous System dataset

## E. Result of Internal Validation

The internal validation measures of connectivity, Dunn index and silhouette width for each of the eight algorithms for the Breast Cancer dataset are shown in *Table* A-1. It

is noticed that hierarchical clustering with two clusters performs the best in the case of connectivity and silhouette width and with four clusters in case of Dunn index. The plots of the connectivity, Dunn index, and silhouette width are given in *Figure* A-11, which indicates that hierarchical clustering outperforms the other clustering algorithms under each validation measure and hence appears to be the method of choice.

**Table A-1:** Scores of Internal Validation Measures for the Breast Cancer dataset

| Clustering Algorithm | Validation Measures | Number of Clusters 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| hierarchical | Connectivity | 2.929 | 5.8579 | 8.7869 | 11.7159 | 14.6448 |
| | Dunn | 0.7 | 0.6464 | 0.779 | 0.6109 | 0.6561 |
| | Silhouette | 0.3772 | 0.3482 | 0.33 | 0.2283 | 0.2021 |
| kmeans | Connectivity | 5.8579 | 16.2127 | 16.546 | 30.0127 | 39.7175 |
| | Dunn | 0.6464 | 0.4386 | 0.4902 | 0.3318 | 0.3538 |
| | Silhouette | 0.374 | 0.2763 | 0.253 | 0.1119 | 0.1019 |
| diana | Connectivity | 5.8579 | 5.8579 | 8.7869 | 19.7528 | 19.9194 |
| | Dunn | 0.6464 | 0.6464 | 0.779 | 0.5083 | 0.5083 |
| | Silhouette | 0.374 | 0.3482 | 0.33 | 0.1891 | 0.1882 |
| som | Connectivity | 2.929 | 17.225 | 42.019 | 48.2409 | 59.0298 |
| | Dunn | 0.7 | 0.3534 | 0.3388 | 0.2798 | 0.3083 |
| | Silhouette | 0.3772 | 0.1222 | 0.0854 | 0.0121 | -0.0197 |
| pam | Connectivity | 28.4897 | 30.6187 | 33.381 | 35.3571 | 37.2361 |
| | Dunn | 0.2716 | 0.2905 | 0.2995 | 0.36 | 0.3837 |
| | Silhouette | 0.0902 | 0.0781 | 0.0852 | 0.0706 | 0.0559 |
| sota | Connectivity | 26.4512 | 35.2464 | 37.7075 | 38.2075 | 41.1448 |
| | Dunn | 0.3005 | 0.2998 | 0.3341 | 0.3341 | 0.3341 |
| | Silhouette | 0.0919 | 0.0873 | 0.0828 | -0.0106 | -0.0429 |
| clara | Connectivity | 28.4897 | 30.6187 | 33.381 | 35.3571 | 37.2361 |
| | Dunn | 0.2716 | 0.2905 | 0.2995 | 0.36 | 0.3837 |
| | Silhouette | 0.0902 | 0.0781 | 0.0852 | 0.0706 | 0.0559 |
| model | Connectivity | 34.3698 | 16.2127 | 16.546 | 25.0968 | 25.7079 |
| | Dunn | 0.3518 | 0.4386 | 0.4902 | 0.4902 | 0.5422 |
| | Silhouette | 0.1651 | 0.2763 | 0.253 | 0.1901 | 0.1787 |

**Optimal Scores:**

| | Score | Method | Clusters |
|---|---|---|---|
| Connectivity | 2.929 | hierarchical | 2 |
| Dunn | 0.779 | hierarchical | 4 |
| Silhouette | 0.3772 | hierarchical | 2 |

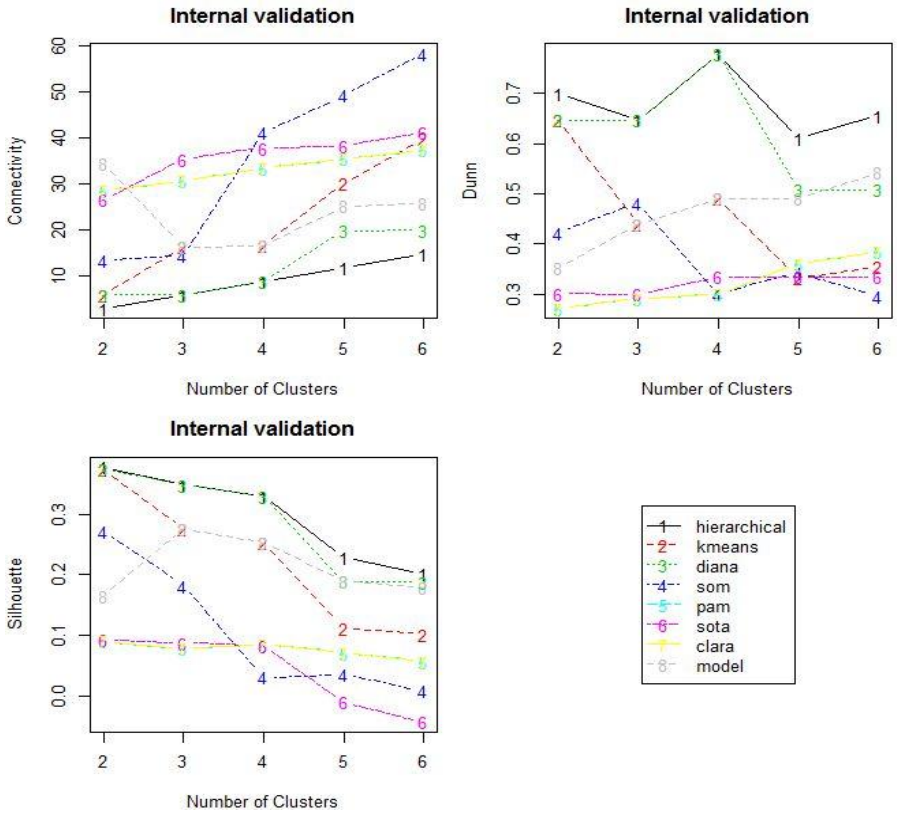



**Figure A-11:** The plots of the connectivity, Dunn index, and silhouette width for the Breast Cancer dataset

The eight algorithms using the Lymphoma dataset are subjected to the calculation of connectivity, Dunn index and silhouette width internal validation measures and the results are shown in *Table* A-2. As in the case in *Table* A-1, it is noticed that hierarchical clustering with two clusters performs the best in the case of connectivity and silhouette width and with three clusters in case of Dunn index. This indicates that hierarchical clustering outperforms the other clustering algorithms under each validation measure and from the plots of connectivity, Dunn index, and silhouette width shown in *Figure* A-12 hierarchical clustering appears to be the method of choice.

**Table A-2:** Scores of Internal Validation Measures for the Lymphoma dataset

| **Clustering Algorithm** | **Validity Measures** | **Number of Clusters** | | | | |
|---|---|---|---|---|---|---|
| | | **2** | **3** | **4** | **5** | **6** |
| hierarchical | Connectivity | 4.2869 | 7.2159 | 9.7159 | 16.9147 | 19.2897 |
| | Dunn | 0.8238 | 0.977 | 0.6611 | 0.4195 | 0.511 |
| | Silhouette | 0.8972 | 0.8406 | 0.8273 | 0.7657 | 0.7259 |
| kmeans | Connectivity | 4.2869 | 13.3524 | 15.8524 | 16.9147 | 19.2897 |
| | Dunn | 0.8238 | 0.2025 | 0.2025 | 0.4195 | 0.511 |
| | Silhouette | 0.8972 | 0.8007 | 0.788 | 0.7657 | 0.7259 |

| | | | | | | |
|---|---|---|---|---|---|---|
| diana | Connectivity | 4.2869 | 7.2159 | 15.6357 | 18.1357 | 19.8607 |
| | Dunn | 0.8238 | 0.977 | 0.2668 | 0.3938 | 0.4111 |
| | Silhouette | 0.8972 | 0.8406 | 0.7847 | 0.7722 | 0.7314 |
| som | Connectivity | 4.2869 | 13.3524 | 15.8798 | 23.9929 | 28.8571 |
| | Dunn | 0.8238 | 0.2025 | 0.0376 | 0.0366 | 0.0331 |
| | Silhouette | 0.8972 | 0.8007 | 0.6156 | 0.5943 | 0.4918 |
| pam | Connectivity | 4.2869 | 13.8044 | 20.5766 | 22.6599 | 25.1599 |
| | Dunn | 0.8238 | 0.029 | 0.0361 | 0.0474 | 0.0699 |
| | Silhouette | 0.8972 | 0.5765 | 0.5986 | 0.6054 | 0.5932 |
| sota | Connectivity | 7.9778 | 11.9647 | 14.4647 | 19.4813 | 20.9813 |
| | Dunn | 0.1042 | 0.2027 | 0.2027 | 0.2395 | 0.4851 |
| | Silhouette | 0.8023 | 0.7918 | 0.7789 | 0.7249 | 0.7347 |
| clara | Connectivity | 4.2869 | 13.3234 | 18.831 | 20.6643 | 25.6242 |
| | Dunn | 0.8238 | 0.029 | 0.0366 | 0.0474 | 0.0652 |
| | Silhouette | 0.8972 | 0.5587 | 0.5964 | 0.6024 | 0.5945 |
| model | Connectivity | 13.9925 | 40.7837 | 39.077 | 45.2472 | 49.2341 |
| | Dunn | 0.0119 | 0.0036 | 0.0034 | 0.0034 | 0.0067 |
| | Silhouette | 0.3321 | -0.0332 | 0.1733 | 0.2006 | 0.24 |

**Optimal Scores**:

| **Measures** | **Score** | **Method** | **Clusters** |
|---|---|---|---|
| Connectivity | 4.2869 | hierarchical | 2 |
| Dunn | 0.977 | hierarchical | 3 |
| Silhouette | 0.8972 | hierarchical | 2 |

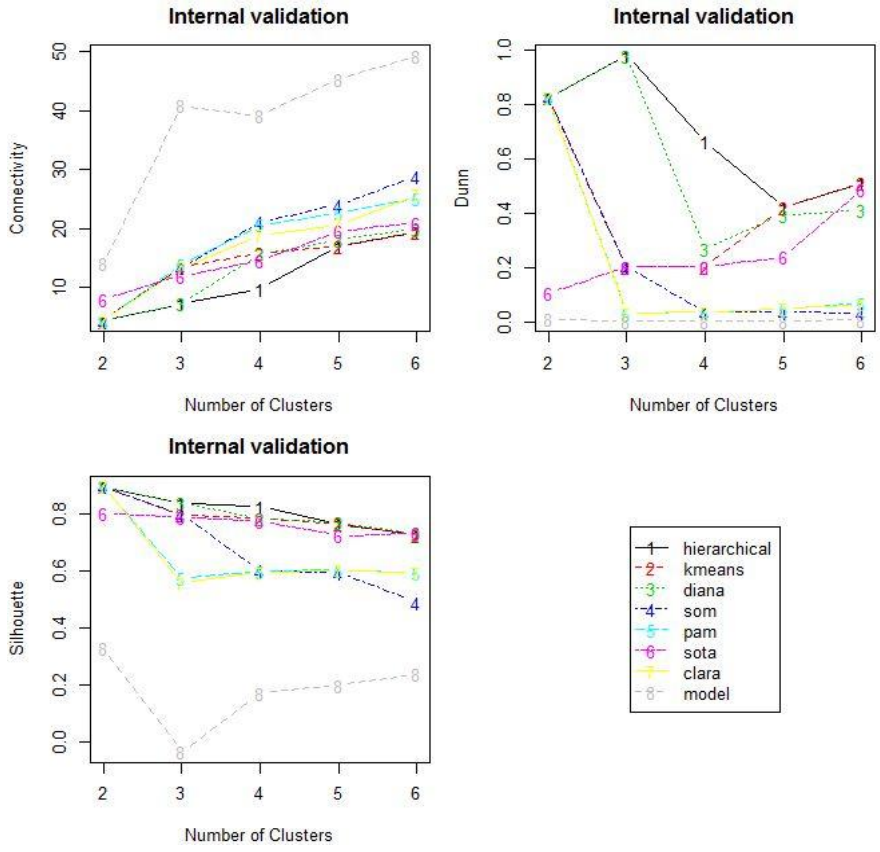



**Figure A-12:** The plots of the connectivity, Dunn index, and silhouette width for the Lymphoma dataset

The internal validation scores for the measures of connectivity, Dunn index and silhouette width for the Embryonal Tumours of Central Nervous System dataset are shown in *Table* A-3. The optimal scores shows that hierarchical clustering with two clusters performs the best in each case, which is confirmed by the plots of the connectivity, Dunn index and silhouette width in *Figure* A-13. SOTA is seen not to perform well as it could not uncover clusters between the ranges four to six.

**Table A-3:** Scores of Internal Validation Measures for the Embryonal Tumours of Central Nervous System dataset

| **Clustering Algorithm** | **Validation Measures** | **Number of Clusters** | | | | |
|---|---|---|---|---|---|---|
| | | **2** | **3** | **4** | **5** | **6** |
| hierarchical | Connectivity | 2.9290 | 5.8579 | 8.7869 | 15.3111 | 18.2401 |
| | Dunn | 1.7906 | 0.6743 | 0.5279 | 0.3530 | 0.4330 |
| | Silhouette | 0.8416 | 0.6274 | 0.5591 | 0.4006 | 0.3225 |
| kmeans | Connectivity | 2.9290 | 15.5933 | 13.9083 | 16.8373 | 19.7663 |
| | Dunn | 1.7906 | 0.2166 | 0.2155 | 0.2622 | 0.2806 |
| | Silhouette | 0.8416 | 0.4468 | 0.3842 | 0.3642 | 0.3405 |
| diana | Connectivity | 2.9290 | 14.2833 | 16.4028 | 18.7984 | 21.7274 |
| | Dunn | 1.7906 | 0.1793 | 0.2814 | 0.3223 | 0.3451 |
| | Silhouette | 0.8416 | 0.4039 | 0.4298 | 0.4193 | 0.3678 |
| som | Connectivity | 2.9290 | 15.5933 | 16.8373 | 28.8837 | 36.7782 |
| | Dunn | 1.7906 | 0.2166 | 0.1653 | 0.0874 | 0.0909 |
| | Silhouette | 0.8416 | 0.4468 | 0.3593 | 0.1655 | 0.1893 |
| pam | Connectivity | 2.9290 | 16.3127 | 17.2472 | 37.7722 | 42.3845 |
| | Dunn | 1.7906 | 0.1589 | 0.2155 | 0.1543 | 0.1543 |
| | Silhouette | 0.8416 | 0.3900 | 0.3876 | 0.2089 | 0.1927 |
| sota | Connectivity | 13.4321 | 14.3667 | NA | NA | NA |
| | Dunn | 0.0601 | 0.0601 | NA | NA | NA |
| | Silhouette | 0.3133 | 0.3036 | NA | NA | NA |
| clara | Connectivity | 2.9290 | 17.4556 | 27.0302 | 39.9345 | 42.3845 |
| | Dunn | 1.7906 | 0.2082 | 0.1146 | 0.1268 | 0.1543 |
| | Silhouette | 0.8416 | 0.4432 | 0.2535 | 0.1889 | 0.1927 |
| model | Connectivity | 21.8623 | 27.5202 | 28.4548 | 38.2885 | 57.5889 |
| | Dunn | 0.0601 | 0.1345 | 0.1345 | 0.2147 | 0.1106 |
| | Silhouette | 0.3055 | 0.2969 | 0.3050 | 0.2924 | 0.0660 |

**Optimal Scores**:

| | Score | Method | Clusters |
|---|---|---|---|
| Connectivity | 2.9290 | hierarchical | 2 |
| Dunn | 1.7906 | hierarchical | 2 |
| Silhouette | 0.8416 | hierarchical | 2 |

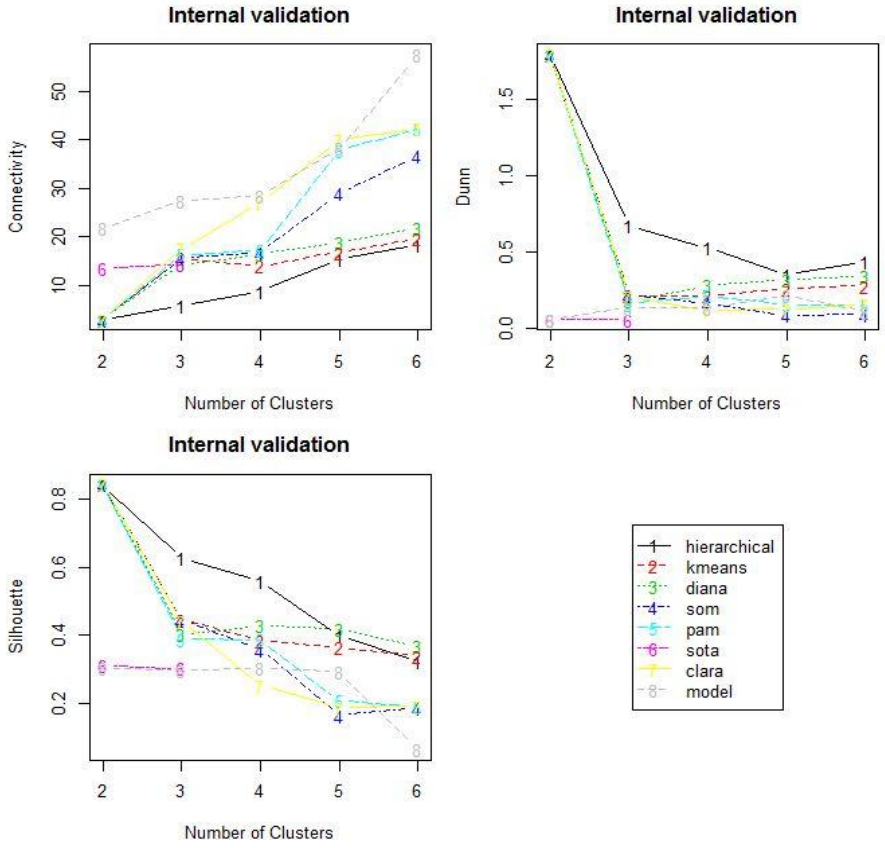



**Figure A-13:**The plots of the connectivity, Dunn index, and silhouette width for the Embryonal Tumours of Central Nervous System dataset

## F. Result of Stability Measures

The results of APN, AD, ADM and FOM for the Breast Cancer dataset are given in *Table* A-4.

**Table A-4:** Scores of Stability Measures for the Breast Cancer dataset

| Clustering Algorithm | Validation Measures | Number of Clusters 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| hierarchical | APN | 0.0022 | 0.0095 | 0 | 0.0007 | 0.0029 |
| | AD | 2.9282 | 2.7599 | 2.5345 | 2.3969 | 2.2656 |
| | ADM | 0.0142 | 0.0721 | 0 | 0.0039 | 0.0131 |
| | FOM | 0.2535 | 0.2455 | 0.2276 | 0.2245 | 0.2188 |
| kmeans | APN | 0.0062 | 0.0189 | 0.0196 | 0.0188 | 0.0614 |
| | AD | 2.8815 | 2.7049 | 2.5194 | 2.3658 | 2.2546 |
| | ADM | 0.0195 | 0.1349 | 0.0772 | 0.0606 | 0.2108 |
| | FOM | 0.2453 | 0.2406 | 0.2297 | 0.2229 | 0.2206 |
| diana | APN | 0.014 | 0.0102 | 0.002 | 0.0155 | 0.0074 |
| | AD | 2.9087 | 2.7618 | 2.5364 | 2.3801 | 2.2426 |
| | ADM | 0.0821 | 0.0772 | 0.0066 | 0.0882 | 0.0459 |
| | FOM | 0.2468 | 0.2464 | 0.2286 | 0.2235 | 0.2182 |
| som | APN | 0.1159 | 0.1977 | 0.2954 | 0.3578 | 0.3588 |
| | AD | 3.0412 | 3.0197 | 2.7384 | 2.6827 | 2.5544 |
| | ADM | 0.4747 | 0.7853 | 0.764 | 0.9293 | 0.9193 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | FOM | 0.248 | 0.2458 | 0.2392 | 0.231 | 0.2326 |
| pam | APN | 0.02 | 0.0161 | 0.0204 | 0.0526 | 0.1105 |
| | AD | 2.9623 | 2.75 | 2.5661 | 2.3923 | 2.3213 |
| | ADM | 0.1316 | 0.0512 | 0.0639 | 0.1034 | 0.2211 |
| | FOM | 0.2572 | 0.2432 | 0.2323 | 0.2228 | 0.2243 |
| sota | APN | 0.1951 | 0.2074 | 0.2061 | 0.1932 | 0.193 |
| | AD | 3.0026 | 2.9005 | 2.8139 | 2.667 | 2.5647 |
| | ADM | 0.4622 | 0.5552 | 0.6112 | 0.6088 | 0.6153 |
| | FOM | 0.257 | 0.2532 | 0.2473 | 0.2388 | 0.2329 |
| clara | APN | 0.02 | 0.0192 | 0.0204 | 0.035 | 0.072 |
| | AD | 2.9623 | 2.7511 | 2.5661 | 2.3873 | 2.297 |
| | ADM | 0.1316 | 0.057 | 0.0639 | 0.0753 | 0.1456 |
| | FOM | 0.2572 | 0.2435 | 0.2323 | 0.2222 | 0.2227 |
| model | APN | 0.0217 | 0.0085 | 0.0226 | 0.0744 | 0.0776 |
| | AD | 2.9038 | 2.6755 | 2.5151 | 2.395 | 2.2251 |
| | ADM | 0.0925 | 0.0363 | 0.0606 | 0.2165 | 0.1665 |
| | FOM | 0.2515 | 0.2364 | 0.2301 | 0.2224 | 0.2114 |

**Optimal Scores:**

| | **Score** | **Method** | **Clusters** |
|---|---|---|---|
| APN | 0 | hierarchical | 4 |
| AD | 2.2251 | model | 6 |
| ADM | 0 | hierarchical | 4 |
| FOM | 0.2114 | model | 6 |

For the APN and ADM measures, values close to zero are preferred. The optimal score in *Table* A-4 shows that hierarchical clustering with four clusters gives the best score, as was also in the case of internal validation. However, for the other two measures model based clustering with six clusters has the best score. It is illustrative to graphically visualize each of the validation measures.

The plots of the APN, AD, and ADM are given in *Figure* A-14. The APN measure shows an interesting trend, in that it initially stabilizes from two to four clusters for all the clustering methods except for SOM and SOTA, but marginally increases afterwards. Though hierarchical clustering with four clusters has the best score, Diana with six clusters is a close second. The AD and FOM measures tend to decrease as the number of clusters increases. Here model based clustering with six clusters has the best overall score, though the other algorithms have similar scores. The plot of the FOM measure is very similar to the AD measure, so it has been omitted from the figure. For the ADM measure hierarchical with four clusters again has the best score.

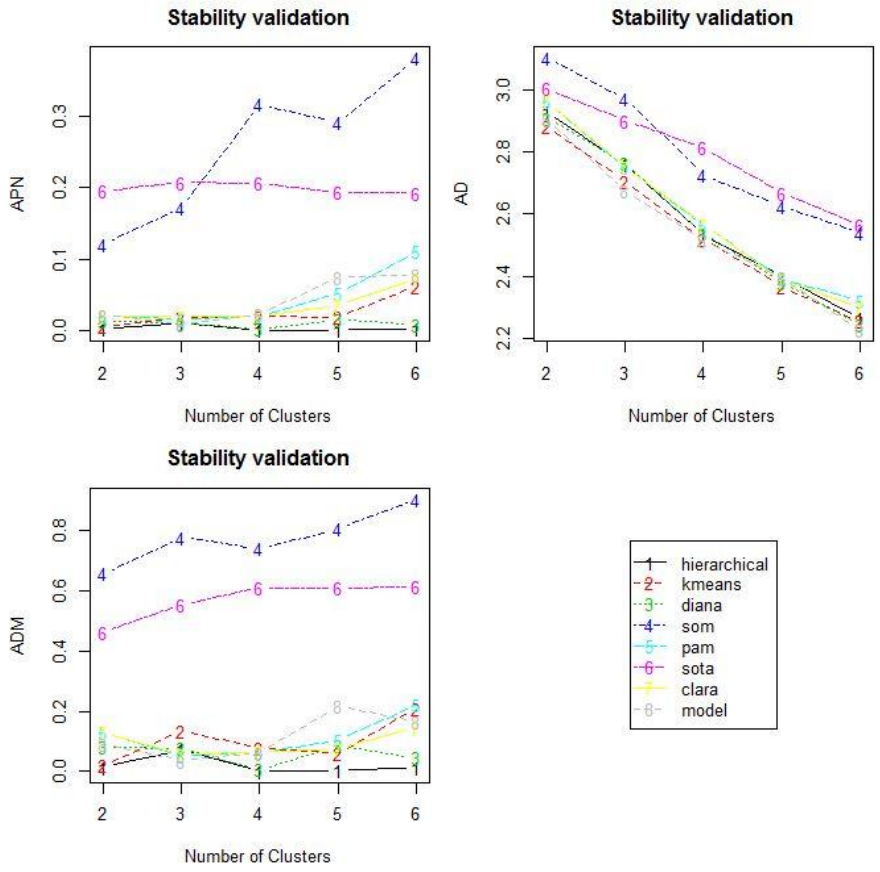



**Figure A-14:**The plots of the APN, AD and ADM of stability measures for the Breast Cancer dataset

For the Lymphoma dataset, the plots of the APN, AD, and ADM are given in *Figure* A-14. Though the graph of APN measure shows DIANA as the most favorable algorithm, one must keep in mind that this algorithm is a special case of the hierarchical algorithm. Thus, hierarchical clustering with two clusters gives the best score and it matches the findings as seen in the case of internal validation in the optimal score given in *Table* A-5. For the AD and FOM measures, PAM with six clusters has the best overall score, but over the entire range of clusters evaluated SOM, K-means, and Diana have comparable performance. Similarly, for the ADM measure hierarchical has a more stable and better performance.

**Table A-5:** Scores of Stability Measures for the Lymphoma dataset

| Clustering Algorithm | Validation Measures | Number of Clusters 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| hierarchical | APN | 0 | 0 | 0.0125 | 0.0065 | 0.0006 |
| | AD | 1530.555 | 1329.008 | 1274.068 | 909.3139 | 809.4373 |
| | ADM | 0 | 0 | 83.4987 | 59.4935 | 2.7571 |
| | FOM | 342.5648 | 291.3158 | 271.6255 | 190.8146 | 157.797 |
| kmeans | APN | 0 | 0.006 | 0.013 | 0 | 0.0006 |
| | AD | 1530.555 | 1093.747 | 1042.737 | 877.4125 | 809.4373 |
| | ADM | 0 | 31.9941 | 86.5842 | 0 | 2.7571 |
| | FOM | 342.5648 | 238.5607 | 215.6405 | 173.3368 | 157.797 |
| diana | APN | 0 | 0 | 0.0009 | 0.0009 | 0.0017 |
| | AD | 1530.555 | 1329.008 | 968.9797 | 900.9588 | 846.0369 |
| | ADM | 0 | 0 | 4.1009 | 4.1009 | 8.9774 |
| | FOM | 342.5648 | 291.3158 | 198.0643 | 179.1709 | 168.7001 |
| som | APN | 0 | 0.0126 | 0.0754 | 0.0216 | 0.0444 |
| | AD | 1530.555 | 1099.501 | 882.7788 | 738.393 | 694.3283 |
| | ADM | 0 | 77.6999 | 203.8568 | 54.9332 | 106.6066 |
| | FOM | 342.5648 | 238.2675 | 214.3498 | 190.1775 | 176.4898 |
| pam | APN | 0 | 0.01 | 0.0221 | 0.0197 | 0.0245 |
| | AD | 1530.555 | 1085.268 | 841.5224 | 722.548 | 657.184 |
| | ADM | 0 | 25.778 | 75.9067 | 58.4361 | 66.9982 |
| | FOM | 342.5648 | 272.8504 | 213.5807 | 176.8744 | 157.5351 |
| sota | APN | 0 | 0 | 0 | 0 | 0 |
| | AD | 1521.36 | 1068.687 | 1000.666 | 915.7482 | 813.0903 |
| | ADM | 0 | 0 | 0 | 0 | 0 |
| | FOM | 439.792 | 232.5206 | 216.3393 | 193.6261 | 156.8454 |
| clara | APN | 0 | 0.0138 | 0.0254 | 0.0354 | 0.0736 |
| | AD | 1530.555 | 1099.382 | 848.081 | 742.9175 | 703.8746 |
| | ADM | 0 | 59.1567 | 64.0555 | 85.9023 | 167.6837 |
| | FOM | 342.5648 | 275.6008 | 209.5931 | 174.8482 | 156.8659 |
| model | APN | 0.0008 | 0.0232 | 0.006 | 0.0137 | 0.0142 |
| | AD | 1988.862 | 1965.832 | 1269.142 | 1192.047 | 739.7668 |
| | ADM | 2.911 | 4.7961 | 1.1664 | 12.7412 | 14.4177 |
| | FOM | 615.289 | 618.8861 | 436.0153 | 434.2291 | 208.9034 |

**Optimal Scores:**

| | Score | Method | Clusters |
|---|---|---|---|
| APN | 0 | hierarchical | 2 |
| AD | 657.184 | pam | 6 |
| ADM | 0 | hierarchical | 2 |
| FOM | 156.8454 | sota | 6 |

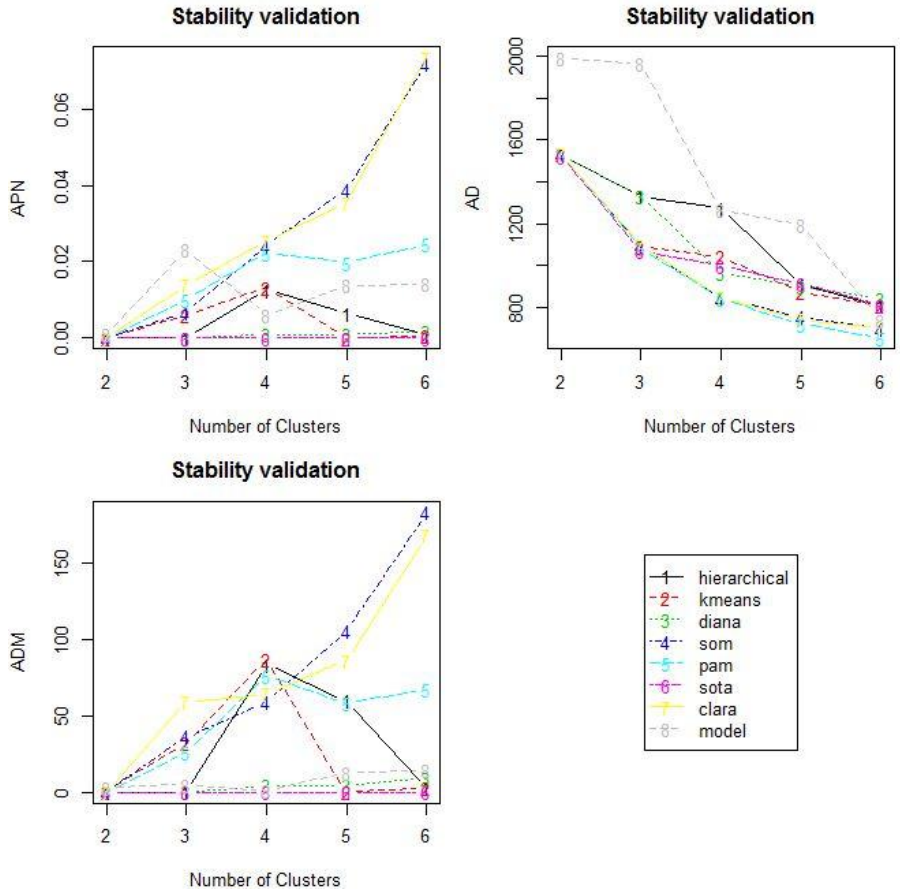



**Figure A-15:**The plots of the APN, AD and ADM of stability measures for the Lymphoma dataset

The plots for the Embryonal Tumours of Central Nervous System dataset of APN, AD, and ADM are given in *Figure* A-16. Hierarchical clustering with two clusters is seen to be performing the best score and it confirms the findings of internal validation, followed by Diana. This is confirmed by the optimal scores given in *Table* A-6. It is seen that PAM and k-means also perform well. For the AD and FOM measures, PAM with six clusters has the best overall score along with k-means. SOTA is seen not to perform well in this case as it could not uncover clusters between the ranges four to six.

**Table A-6:** Scores of Stability Measures for the Embryonal Tumours of Central Nervous System dataset

| Clustering Algorithm | Validation Measures | Number of Clusters 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| hierarchical | APN | 0 | 0 | 0.017 | 0.0088 | 0.0079 |
| | AD | 2831.785 | 2575.523 | 2378.493 | 2153.259 | 2021.896 |
| | ADM | 0 | 0 | 81.2571 | 102.1255 | 74.6602 |
| | FOM | 360.8386 | 332.3272 | 298.2834 | 278.2367 | 266.9793 |
| kmeans | APN | 0 | 0.0204 | 0.0069 | 0.039 | 0.0174 |
| | AD | 2831.785 | 2328.474 | 2027.98 | 1900.727 | 1776.726 |
| | ADM | 0 | 72.412 | 35.2552 | 108.3335 | 62.6617 |
| | FOM | 360.8386 | 295.5827 | 263.1819 | 236.2135 | 222.3414 |
| diana | APN | 0 | 0.0039 | 0 | 0.0019 | 0.0046 |
| | AD | 2831.785 | 2405.219 | 2088.549 | 1970.159 | 1859.723 |
| | ADM | 0 | 107.0119 | 0 | 13.6494 | 31.417 |
| | FOM | 360.8386 | 311.1003 | 263.5247 | 259.23 | 249.3723 |
| som | APN | 0.095 | 0.0965 | 0.1939 | 0.2826 | 0.2235 |
| | AD | 3010.729 | 2446.893 | 2219.343 | 2144.637 | 1928.805 |
| | ADM | 418.846 | 495.1717 | 582.3461 | 733.2932 | 601.8914 |
| | FOM | 426.7409 | 359.6395 | 309.9975 | 296.1228 | 277.789 |
| pam | APN | 0 | 0.0291 | 0.0327 | 0.0264 | 0.0278 |
| | AD | 2831.785 | 2303.267 | 2068.245 | 1884.061 | 1762.742 |
| | ADM | 0 | 136.7734 | 128.434 | 96.6943 | 92.9114 |
| | FOM | 360.8386 | 302.1433 | 270.1488 | 252.8933 | 232.0985 |
| sota | APN | 0.0798 | 0.1756 | NA | NA | NA |
| | AD | 3090.101 | 2973.923 | NA | NA | NA |
| | ADM | 262.433 | 515.4014 | NA | NA | NA |
| | FOM | 413.3749 | 392.2832 | NA | NA | NA |
| clara | APN | 0 | 0.0832 | 0.1688 | 0.0691 | 0.0947 |
| | AD | 2831.785 | 2368.853 | 2248.118 | 1905.098 | 1791.091 |
| | ADM | 0 | 294.6688 | 657.8768 | 164.1097 | 200.0441 |
| | FOM | 360.8386 | 300.8604 | 279.2093 | 256.7445 | 232.4297 |
| model | APN | 0.0095 | 0.0234 | 0.0408 | 0.0809 | 0.0891 |
| | AD | 3077.805 | 2366.015 | 2141.536 | 1936.032 | 1856.468 |
| | ADM | 30.7028 | 64.4391 | 101.093 | 234.2042 | 278.5476 |
| | FOM | 430.198 | 328.2964 | 304.3724 | 269.9173 | 258.5611 |

**Optimal Scores:**

| | Score | Method | Clusters |
|---|---|---|---|
| APN | 0 | hierarchical | 2 |
| AD | 1762.742 | pam | 6 |
| ADM | 0 | hierarchical | 2 |
| FOM | 222.3414 | kmeans | 6 |

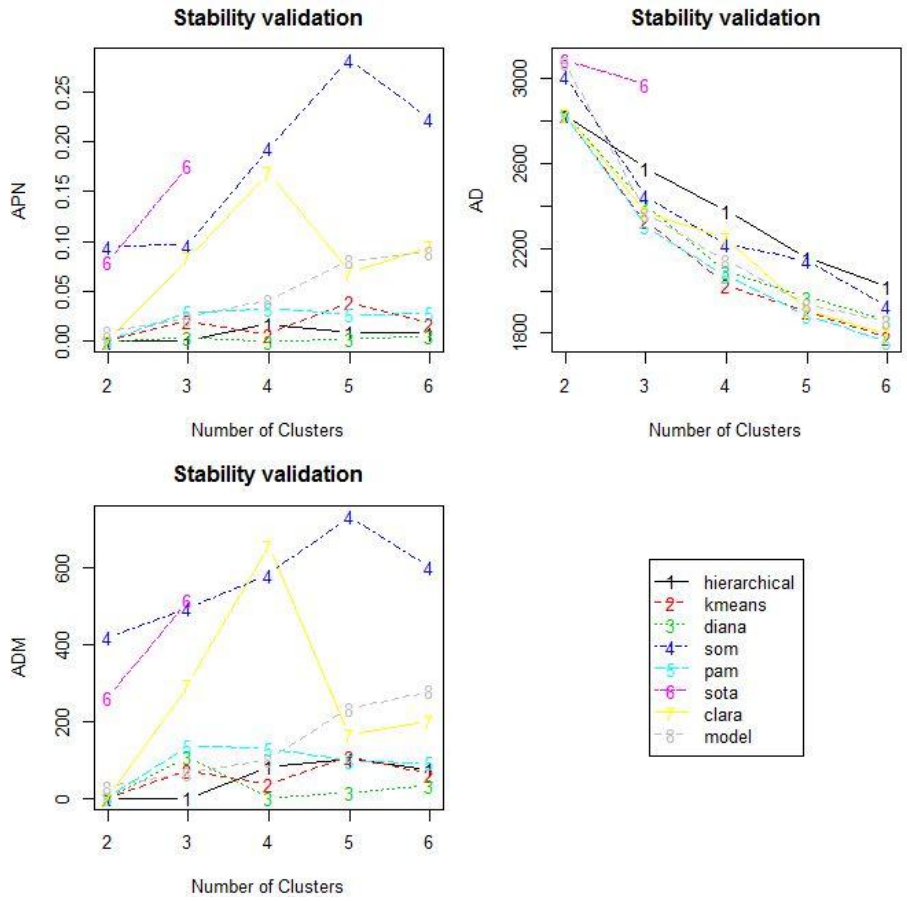



**Figure A-16:**The plots of the APN, AD and ADM of stability measures for the Embryonal Tumours of Central Nervous System dataset

## G. Results of BHI and BSI

The BHI and the BSI values were computed for each clustering algorithm in the range of cluster numbers from two to six. The breast cancer data is considered first. *Table* A-7 shows the scores for the Breast Cancer dataset and it is seen that DIANA has the highest BHI score for six clusters and the highest BSI score is by hierarchical algorithm for two clusters, which indicates that consistency of clustering for genes with similar biological functionality is given by hierarchical algorithm.

**Table A-7:** Scores of BHI and BSI for the Breast Cancer dataset

| **Algorithm** | **Measure** | **Number of Clusters** | | | | |
|---|---|---|---|---|---|---|
| | | **2** | **3** | **4** | **5** | **6** |
| hierarchical | BHI | 0.3172 | 0.3095 | 0.3011 | 0.2966 | 0.2882 |
| | BSI | 0.8253 | 0.6611 | 0.5725 | 0.5329 | 0.5179 |
| kmeans | BHI | 0.4047 | 0.3639 | 0.3889 | 0.3529 | 0.2879 |
| | BSI | 0.6371 | 0.4925 | 0.4579 | 0.3006 | 0.2511 |
| diana | BHI | 0.4047 | 0.3095 | 0.3011 | 0.3861 | 0.4241 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | BSI | 0.6574 | 0.6605 | 0.5708 | 0.3361 | 0.3021 |
| som | BHI | 0.3163 | 0.3156 | 0.2947 | 0.2989 | 0.2361 |
| | BSI | 0.7014 | 0.5614 | 0.3147 | 0.2431 | 0.1773 |
| pam | BHI | 0.3052 | 0.3022 | 0.2979 | 0.2853 | 0.2839 |
| | BSI | 0.5904 | 0.5511 | 0.4181 | 0.3831 | 0.3612 |
| sota | BHI | 0.3041 | 0.2833 | 0.2125 | 0.1708 | 0.1867 |
| | BSI | 0.4697 | 0.4221 | 0.4131 | 0.4092 | 0.3934 |
| clara | BHI | 0.3052 | 0.3022 | 0.2979 | 0.2853 | 0.2839 |
| | BSI | 0.5904 | 0.5501 | 0.4181 | 0.3895 | 0.3725 |
| model | BHI | 0.3091 | 0.3639 | 0.3889 | 0.2651 | 0.2651 |
| | BSI | 0.3612 | 0.4574 | 0.4493 | 0.3891 | 0.3246 |

**Optimal Scores:**

| | **Score** | **Method** | **Clusters** |
|---|---|---|---|
| BHI | 0.424 | diana | 6 |
| BSI | 0.8253 | hierarchical | 2 |

*Figure* A-17 shows the plots of BHI for the eight clustering algorithms which reveal that DIANA happens to produce most homogeneous biological clusters based on this dataset and the results are statistically significant when the number of clusters is between four and six.

The plots of BSI are shown in *Figure* A-18 and hierarchical algorithm seems to be the most stable in its capability of producing clusters using reduced datasets that are biologically alike. Considering both indices, it can be said that hierarchical algorithm is the best choice for this dataset to maximize the biological homogeneity and DIANA can be a worthwhile consideration if six clusters are desired.

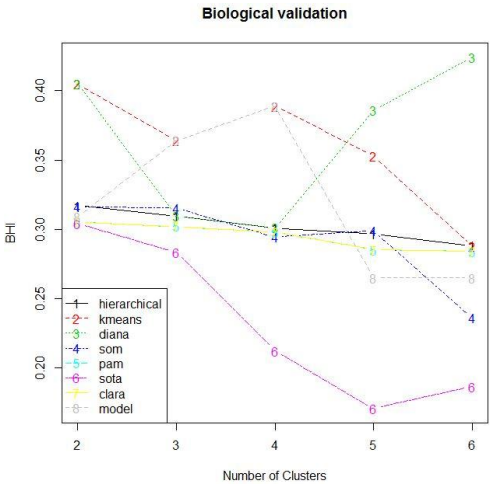



**Figure A-17:**BHI plot for Breast Cancer dataset

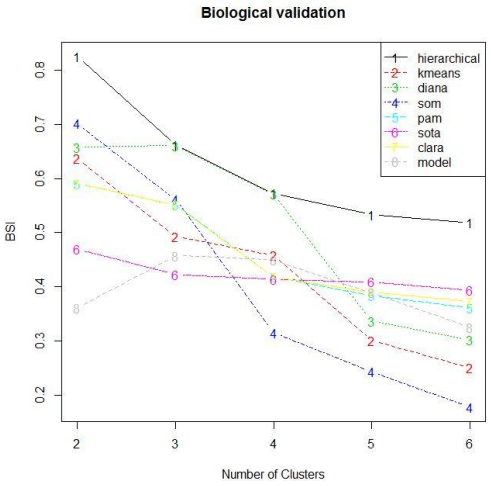



**Figure A-18:**BSI plot for Breast Cancer dataset

The scores for the Lymphoma dataset is shown in *Table* A-8 and it is seen that SOM has the highest BHI score for six clusters and the highest BSI score is again by hierarchical algorithm for two clusters, which indicates its consistency to produce most homogeneous biological clusters based on this dataset.

**Table A-8:** Scores of BHI and BSI for the Lymphoma dataset

| **Algorithm** | **Measure** | **Number of Clusters** | | | | |
|---|---|---|---|---|---|---|
| | | **2** | **3** | **4** | **5** | **6** |
| hierarchical | BHI | 0.1827 | 0.1827 | 0.1827 | 0.1899 | 0.1711 |
| | BSI | 0.9704 | 0.9704 | 0.9593 | 0.8566 | 0.8425 |
| kmeans | BHI | 0.1827 | 0.1329 | 0.1329 | 0.1899 | 0.1711 |
| | BSI | 0.9704 | 0.8585 | 0.8554 | 0.8433 | 0.8425 |
| diana | BHI | 0.1827 | 0.1827 | 0.1329 | 0.1329 | 0.0608 |
| | BSI | 0.9704 | 0.9704 | 0.8632 | 0.8632 | 0.8619 |
| som | BHI | 0.1827 | 0.1329 | 0.1783 | 0.1728 | 0.1977 |
| | BSI | 0.9704 | 0.8570 | 0.5474 | 0.5077 | 0.3879 |
| pam | BHI | 0.1827 | 0.1901 | 0.1813 | 0.1813 | 0.1813 |
| | BSI | 0.9704 | 0.5507 | 0.5242 | 0.5239 | 0.5218 |
| sota | BHI | 0.1566 | 0.1899 | 0.1899 | 0.1711 | 0.1711 |
| | BSI | 0.8433 | 0.8433 | 0.8433 | 0.8425 | 0.8425 |
| clara | BHI | 0.1827 | 0.1901 | 0.1284 | 0.1284 | 0.1284 |
| | BSI | 0.9704 | 0.5544 | 0.5207 | 0.5187 | 0.4993 |
| model | BHI | 0.1821 | 0.1775 | 0.1701 | 0.1717 | 0.1851 |
| | BSI | 0.4830 | 0.2983 | 0.2476 | 0.2102 | 0.2098 |

**Optimal Scores:**

| | **Score** | **Method** | **Clusters** |
|---|---|---|---|
| **BHI** | 0.1977 | som | 6 |
| **BSI** | 0.9704 | hierarchical | 2 |

*Figure* A-19 shows that SOM produces the most homogeneous biological clusters when six clusters are required and hierarchical is the most consistent of all the algorithms. The plots of BSI are shown in *Figure* A-20 and hierarchical algorithm appears to be the most stable in its capability of producing clusters that are biologically alike and model-based clustering appears to be the least stable. It can be concluded that hierarchical algorithm seems to be the best choice for this dataset to maximize the biological homogeneity, considering both the indices.

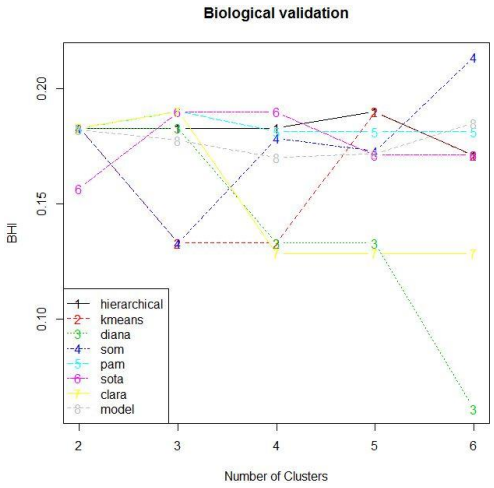



**Figure A-19:** BHI plot for Lymphoma dataset

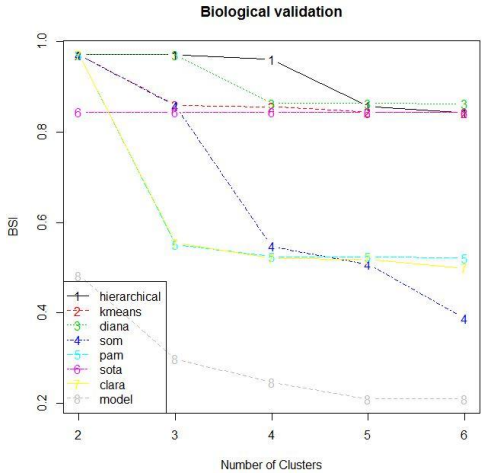



**Figure A-20:** BSI plot for Lymphoma dataset

Finally, for the Embryonal Tumours of Central Nervous System dataset, the BHI and the BSI scores are shown in *Table* A-9 and in both cases, hierarchical scores the highest points for producing biological significant clusters.

**Table A-9:** Scores of BHI and BSI for the CNS dataset

| **Algorithms** | **Measure** | **Number of Clusters** | | | | |
|---|---|---|---|---|---|---|
| | | **2** | **3** | **4** | **5** | **6** |
| hierarchical | BHI | 0.2604 | 0.2556 | 0.2551 | 0.3725 | 0.375 |
| | BSI | 0.8654 | 0.8153 | 0.7806 | 0.6187 | 0.5784 |
| kmeans | BHI | 0.2604 | 0.3004 | 0.267 | 0.2685 | 0.2712 |
| | BSI | 0.8654 | 0.5328 | 0.4299 | 0.407 | 0.396 |
| diana | BHI | 0.2604 | 0.2791 | 0.2766 | 0.275 | 0.2771 |
| | BSI | 0.8654 | 0.6094 | 0.5388 | 0.5323 | 0.5036 |
| som | BHI | 0.2604 | 0.3004 | 0.2873 | 0.2715 | 0.324 |
| | BSI | 0.7783 | 0.5109 | 0.3581 | 0.2972 | 0.2376 |
| pam | BHI | 0.2604 | 0.2743 | 0.2656 | 0.3019 | 0.2993 |
| | BSI | 0.8654 | 0.4532 | 0.4405 | 0.299 | 0.2941 |
| sota | BHI | 0.2813 | 0.2747 | NA | NA | NA |
| | BSI | 0.3908 | 0.3515 | NA | NA | NA |
| clara | BHI | 0.2604 | 0.3077 | 0.2572 | 0.2977 | 0.2993 |
| | BSI | 0.8654 | 0.5033 | 0.43 | 0.293 | 0.2836 |
| model | BHI | 0.2687 | 0.269 | 0.2646 | 0.2783 | 0.2635 |
| | BSI | 0.376 | 0.3659 | 0.3472 | 0.2918 | 0.2192 |

**Optimal Scores:**

| | **Score** | **Method** | **Clusters** |
|---|---|---|---|
| **BHI** | 0.375 | hierarchical | 6 |
| **BSI** | 0.8654 | hierarchical | 2 |

Although hierarchical shows a marked increase for cluster sizes of five or six, SOM can be the algorithm of choice when four clusters are desired, as shown in *Figure* A-

21. When the plots of BSI as shown in *Figure* A-22 are compared, it can be seen that all the clustering algorithms have produced significantly consistent results barring SOTA, as it could not generate clusters between the ranges four to six. Hierarchical algorithm seems to be the most stable in its ability of producing biologically relevant clusters and on comparing both the indices, it can be concluded that hierarchical algorithm is the best choice for this dataset to maximize the biological homogeneity.

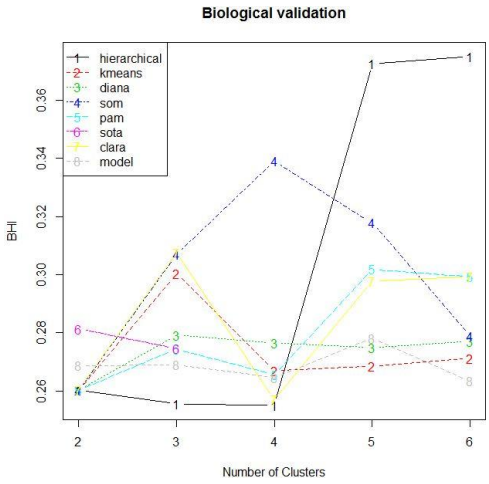



**Figure A-21:**BHI plot for CNS dataset

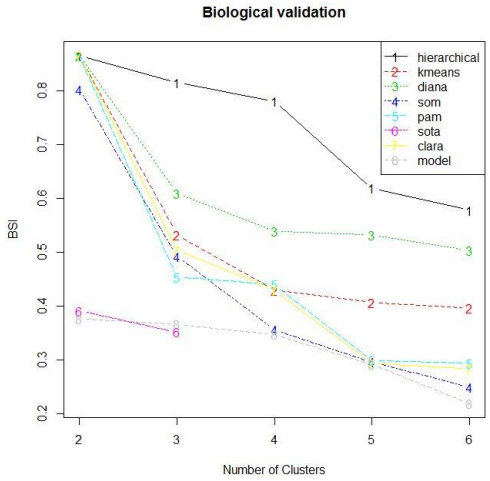



**Figure A-22:**BSI plot for CNS dataset