Bioinformation **Volume 5** *open access*
www.bioinformation.net Issue 6 **Hypothesis**

Supplementary material:

Table 1: List of the computer programs with their URLs.

Program	Version	URL
Decrease Redundancy	N/A	http://www.expasy.org/tools/redundancy
cd-hit	3.1.1	http://www.bioinformatics.org/cd-hit
Pisces	N/A	http://dunbrack.fccc.edu/PISCES.php
BlastClust (BLAST)	2.2.16	http://blast.ncbi.nlm.nih.gov
SkipRedundant (EMBOSS)	6.2.0	http://emboss.sourceforge.net

Table 2: Summary information on the datasets (nres indicates the number of residues).

Dataset name	Dataset content
D_100_0	100 sequences with nres ≤ 100
D_100_100	100 sequences with 100 < nres ≤ 200
D_100_200	100 sequences with 200 < nres ≤ 300
D_100_300	100 sequences with 300 < nres ≤ 400
D_100_400	100 sequences with 400 < nres ≤ 500
D_100_500	100 sequences with 500 < nres ≤ 600
D_100_600	100 sequences with 600 < nres ≤ 700
D_100_700	100 sequences with 700 < nres ≤ 800
D_100_800	100 sequences with 800 < nres ≤ 900
D_100_900	100 sequences with 900 < nres ≤ 1000
D_100_1000	100 sequences with nres > 1000
D_1000_0	1000 sequences with nres ≤ 100
D_1000_100	1000 sequences with 100 < nres ≤ 200
D_1000_200	1000 sequences with 200 < nres ≤ 300
D_1000_300	1000 sequences with 300 < nres ≤ 400
D_1000_400	1000 sequences with 400 < nres ≤ 500
D_1000_500	1000 sequences with 500 < nres ≤ 600
D_1000_600	1000 sequences with 600 < nres ≤ 700
D_1000_700	1000 sequences with 700 < nres ≤ 800
D_1000_800	1000 sequences with 800 < nres ≤ 900
D_1000_900	1000 sequences with 900 < nres ≤ 1000
D_1000_1000	1000 sequences with nres > 1000
D_10000_0	10000 sequences with nres ≤ 100
D_10000_100	10000 sequences with 100 < nres ≤ 200
D_10000_200	10000 sequences with 200 < nres ≤ 300
D_10000_300	10000 sequences with 300 < nres ≤ 400
D_10000_400	10000 sequences with 400 < nres ≤ 500
D_10000_500	10000 sequences with 500 < nres ≤ 600
D_10000_600	10000 sequences with 600 < nres ≤ 700
D_10000_700	10000 sequences with 700 < nres ≤ 800
D_10000_800	10000 sequences with 800 < nres ≤ 900
D_10000_900	10000 sequences with 900 < nres ≤ 1000
D_10000_1000	10000 sequences with nres > 1000

Table 3: The main features of the different computer programs that were used.

Program name	Stand alone (OS)	% sequence identity threshold	Output is dependent on the input order	Output format
Decrease Redundancy	No	0 – 100 (any value)	Yes	FASTA
cd-hit	Yes (Linux, Windows)	40 – 100 (any value)	Yes	FASTA
Pisces	Yes (Linux)	5 - 100 (any value)	No	List of identification codes
BlastClust	Yes (Linux, Windows)	0 – 100 (any value)	No	List of identification codes
SkipRedundant	Yes (Linux)	0 – 100 (any value)	Yes	FASTA

Table 4: The percentage of sequences found in the output relative to the input (Ptot). The thresholds of percentage of sequence identity are indicated as 'Max PID'. These data are the averages of the results obtained with all datasets.

Program	**Max PID 40%**	**Max PID 50%**	**Max PID 75%**	**Max PID 90%**
Decrease redundancy	Ptot = 95%	Ptot = 96%	Ptot = 97 %	Ptot = 99%
cd-hit	Ptot = 88%	Ptot = 91%	Ptot = 94%	Ptot = 98%
Pisces	Ptot = 88%	Ptot = 91%	Ptot = 95 %	Ptot = 98 %
BlastClust	Ptot = 89 %	Ptot = 91%	Ptot = 94%	Ptot = 98 %
SkipRedundant	Ptot = - *%	Ptot = -*%	Ptot = 60%	Ptot = 68%

* Despite several attempts, the program reported results containing few clusters which were not further analyzed.

Table 5: Average overlap (standard deviation) between the outputs of different programs observed at different thresholds of sequence identity - Max PID. Averages and standard deviations were computed by using all the data sets.

Max PID 40%	**Decrease redundancy**	**cd-hit**	**Pisces**	**BlastClust**	**Skip Redundant**
Decrease redundancy	100%	88% (6)	89% (5)	89% (5)	N/A
cd-hit		100%	95% (2)	95% (3)	N/A
Pisces			100%	95% (3)	N/A
BlastClust				100%	N/A
Skip Redundant	N/A	N/A	N/A	N/A	N/A
Max PID 50%	**Decrease redundancy**	**cd-hit**	**Pisces**	**BlastClust**	**Skip Redundant**
Decrease redundancy	100%	89% (5)	90% (5)	90% (3)	N/A
cd-hit		100%	96% (2)	94% (1)	N/A
Pisces			100%	97% (1)	N/A
BlastClust				100%	N/A
Skip Redundant					N/A
Max PID 75%	**Decrease redundancy**	**cd-hit**	**Pisces**	**BlastClust**	**Skip Redundant**
Decrease redundancy	100%	90% (4)	90% (4)	92% (4)	91% (5)
cd-hit		100%	96% (1)	95% (2)	99% (1)
Pisces			100%	98% (1)	97% (2)
BlastClust				100%	97% (2)
Skip Redundant					100%
Max PID 90%	**Decrease redundancy**	**cd-hit**	**Pisces**	**BlastClust**	**Skip Redundant**
Decrease redundancy	100%	92% (4)	92% (4)	95% (3)	94% (4)
cd-hit		100%	95% (1)	96% (1)	99% (1)
Pisces			100%	98% (1)	97% (1)
BlastClust				100%	97% (1)
Skip Redundant					100%