

**Roman Matuszewski**

University of Bialystok  
Bialystok, Poland

**Hanna Sovalat**

Institut de Recherche en Hématologie  
et Transplantations, Mulhouse, France

## TAXONOMICAL CLASSIFICATION OF HEMATOPOIETIC CD34+ CELL SUBSETS FROM DIVERSE ORIGIN<sup>1</sup>

**Abstract:** Taxonomy is one of the numerous methods of classifying empirical data. In this work the *Wroclaw Taxonomy* has been chosen (this method has been previously implemented in the classification of bacteria *yersinia pestis*) [1, 2]. The aim is to group 5 different sources of HPC ( $j = 5$  of the studied objects): *normal* and *mobilized BM*, *normal PB*, *leukapheresis products*, *cord blood*, depending on 13 common features which describe the objects (i.e.  $i = 13$  measurement variables). The objects will be compared between themselves but not in reference to the optimal model.

**Acknowledgements:** Sections of the work by Roman Matuszewski have been supported by the Polish State Scientific Research Committee grant 3 T11F 011 30 “*Temporal representation of the knowledge and their implementation in the medical systems*”. We also thank prof. Andrzej Trybulec for his valuable suggestions.

Taxonomy is the practice and science of classification. Originally, the term taxonomy referred to the science of classifying living organisms; however, the term is now applied in a wider, more general sense and now may refer to the *classification* of objects, as well as to the *principles* underlying such classification.

Taxonomies, or taxonomic schemes, are composed of taxonomic units or kinds of objects that are arranged frequently in a hierarchical structure. A taxonomy might also be a simple organization of kinds of objects into groups. Mathematically, a hierarchical taxonomy is a tree structure of classifications for a given set of objects.

The empirical data has been taken from the paper [3]. We have two series of data:

---

<sup>1</sup> This work was presented in the poster session at the International Symposium *Bio-engineering and Regenerative Medicine*, Sept. 24–26, 2007, Mulhouse, France.

TABLE 1.

Frequency of *early cell subsets* and *cellular adhesion molecules expression* within the CD34+ cell population derived from five sources of Hematopoietic Progenitor Cells – **HPC** (% of total CD34+ cells).

**nBM** = normal bone marrow; **mBM** = mobilized bone marrow;

**nPB** = normal peripheral blood; **LKP** = leukapheresis product;

**CB** = cord blood.

It creates the table  $\mathbf{a}_{i,j}$  of 65 empirical measurements:

	j→	1	2	3	4	5
i ↓	CD34+ cell subsets	nBM (n=13) (steady state)	mBM (n=16)	nPB (n=13) (steady state)	LKP (n=29)	CB (n=20)
1	CD38 <sup>-</sup>	3.47 ± 0.06	3.51 ± 4.40	1.40 ± 1.87	2.10 ± 0.90	15.59 ± 7.84
2	HLA-DR <sup>-</sup>	4.77 ± 1.60	2.91 ± 0.50	5.71 ± 4.15	0.80 ± 0.58	8.21 ± 3.99
3	CD90 <sup>+</sup>	13.28 ± 8.60	12.96 ± 7.83	ND	19.10 ± 12.89	ND
4	CD117 <sup>+</sup>	11.12 ± 9.91	49.20 ± 6.20	8.12 ± 4.58	61.48 ± 8.90	16.31 ± 4.80
5	PgP170 <sup>+</sup>	2.06 ± 0.78	3.03 ± 0.20	2.33 ± 0.16	3.46 ± 1.62	5.49 ± 2.54
6	CD11a+	64.54+25.59	88.55+ 3.77	75.54+11.08	90.88+ 4.48	89.43+ 8.57
7	CD11b+	8.27+ 2.40	4.17+ 2.02	3.99+ 1.85	3.75+ 1.52	2.87+ 0.35
8	CD49d+	94.24+ 9.64	97.00+ 5.10	ND	68.45+ 9.96	88.74+ 7.85
9	CD49e+	48.75+ 7.88	45.70+ 5.87	ND	38.90+ 5.42	ND
10	CD54+	11.29+16.12	26.27+ 5.16	14.89+ 9.70	50.80+14.80	24.89+11.03
11	CD58+	67.73+20.60	94.40+ 9.74	75.82+11.66	100	74.85+19.80
12	CD44+	90.61+11.49	89.00+11.30	85.96+ 6.83	100	97.27+ 2.25
13	CD62L+	53.13+12.02	51.37+15.80	57.86+15.26	78.91+ 6.19	73.47+14.09

The data express the mean ± standard deviation. ND = not determined due to cell subset barely detectable.

TABLE 2.

Antigen density of *early marker* and *cellular adhesion molecules* on CD34+ cells derived from five sources of Hematopoietic Progenitor Cells – **HPC** (Nb. of ABC x 10<sup>3</sup> molecules / cell).

**nBM** = normal bone marrow; **mBM** = mobilized bone marrow;

**nPB** = normal peripheral blood; **LKP** = leukapheresis product;

**CB** = cord blood.

*Taxonomical classification of hematopoietic CD34+ cell subsets...*

It creates the table  $\mathbf{a}_{i,j}$  of 65 empirical measurements:

	$j \rightarrow$	1	2	3	4	5
$i$ $\downarrow$	CD34+ cell subsets	nBM (n=13) In steady state	mBM (n=16)	nPb (n=13) In steady state	LKP (n=29)	CB (n=20)
1	CD38 <sup>+</sup>	56.60 ± 15.50	48.97 ± 18.31	35.00 ± 9.80	31.30 ± 16.00	37.00 ± 16.20
2	HLA-DR <sup>+</sup>	196.25 ± 8.92	135.42 ± 25.62	89.16 ± 12.14	79.06 ± 8.13	57.28 ± 21.30
3	CD90 (Thy-1) <sup>+</sup>	11.04 ± 5.50	11.00 ± 6.18	ND	9.40 ± 1.71	ND
4	CD117 (c-kit) <sup>+</sup>	16.74 ± 3.48	8.62 ± 3.12	9.01 ± 4.20	4.96 ± 1.42	11.99 ± 1.43
5	PgP170 <sup>+</sup>	18.62 ± 6.00	18.05 ± 1.12	17.38 ± 2.65	17.76 ± 4.26	36.83 ± 2.34
	Integrin family					
6	CD11a	31.88+ 6.19	22.49+ 8.03	18.27+ 5.15	9.44+ 2.35	14.34+ 9.30
7	CD11b	13.15+ 5.75	11.79+ 4.74	13.04+ 8.45	10.20+ 9.35	12.56+ 1.30
8	CD49d	26.61+ 9.76	17.66+ 3.55	ND	8.23+ 1.39	23.60+ 5.13
9	CD49e	30.15+ 8.42	33.62+ 9.30	ND	18.32+ 4.71	ND
	Ig super family					
10	CD54	10.03+ 1.81	7.75+ 4.32	9.10+ 4.26	3.98+ 1.11	8.49+ 5.72
11	CD58	17.25+ 7.08	20.98+13.69	15.75+ 6.91	9.49+ 2.03	18.41+ 9.04
	Homing associated family					
12	CD44	144.89+24.09	124.93+24.89	103.23+28.32	96.58+17.77	160.36+42.64
13	CD62L	25.09+13.72	24.65+ 7.74	15.31+ 8.67	8.87+ 4.33	60.04+29.10

The data express the mean ± standard deviation. ND = not determined due to cell subset barely detectable.

The aim is to group 5 different sources of **HPC** in every series of data:  $j = 5$  of the studied objects, depending on 13 common features which describe the objects:  $i = 13$  measurement variables. The 5 objects will be compared between themselves but not in reference to the optimal model.

The concept enabling reciprocal comparison of objects through the analysis of their variables is similarity. The similarity of objects in taxonomy is measured by the means of distance. If the distance is smaller, objects differ between themselves less (they are closer to each other). The matrix  $\mathbf{a}_{i,j}$  of empirical data consisting of 65 measurements has been standardized through the arithmetic mean for every variable:  $s_{i,j} = \frac{a_{i,j}}{\bar{a}_i}$ .

The *city metric* (Manhattan, Hamming distance) has been implemented which gives similar results as the Euclidean metric. Smaller is here, however, the influence of singular large detached differences. Properties of the metric  $\mathbf{d}$ :

1. identity of indiscernibles:  $d_{m,n} = 0$  iff  $m = n$ ,
2. symmetry:  $d_{m,n} = d_{n,m}$ ,
3. triangle inequality:  $d_{m,n} \leq d_{m,p} + d_{p,n}$ .

Matrix of distances  $d_{m,n} = \sum_{k=1}^i |s_{k,m} - s_{k,n}|$ , where:  $m, n = 1, \dots, j$ ,  $d_{m,n}$  – the distance between a pair of objects  $m$  and  $n$ , is a synthetic measurement of all variables.

Results for TABLE 1.

Matrix of distances  $\mathbf{d}_{m,n}$

$j \rightarrow$	1	2	3	4	5
$\downarrow$					
1	0	4,33337	6,85084	8,41278	10,1635
2	4,33337	0	8,18558	4,33525	9,66091
3	6,85084	8,18558	0	10,3664	6,98268
4	8,41278	4,33525	10,3664	0	11,9020
5	10,1635	9,66091	6,98268	11,9020	0

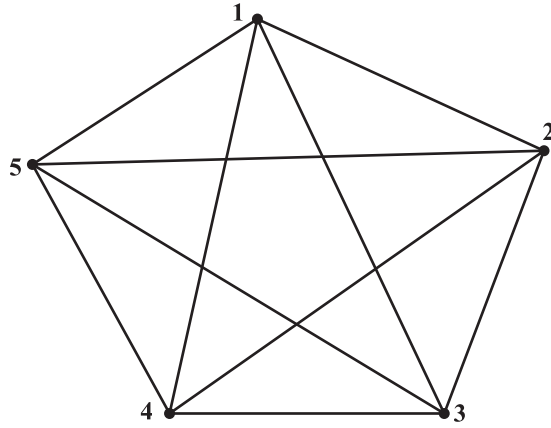
This gives us a set of 10 distances  $\frac{j!}{r!(j-r)!} = 10$ , where  $j = 5$ ,  $r = 2$  (pair).

In arithmetic order

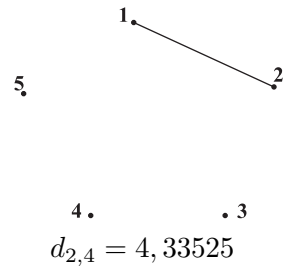
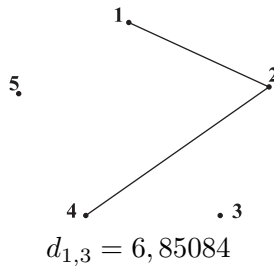
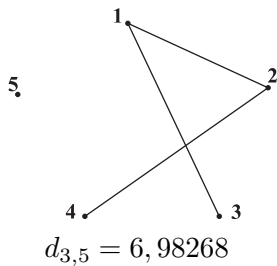
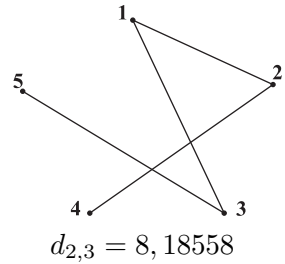
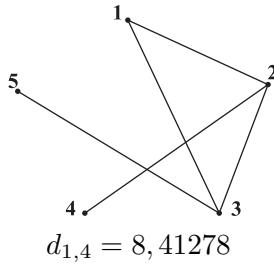
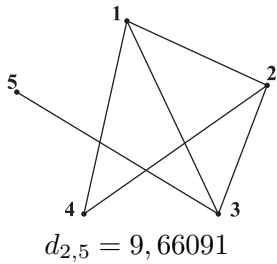
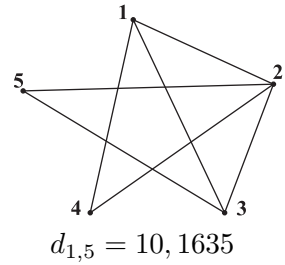
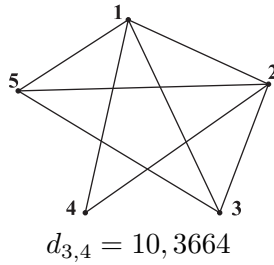
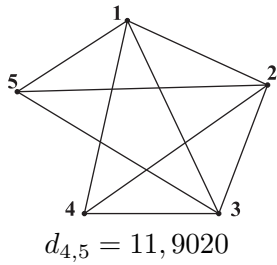
$$\begin{aligned}
 d_{1,2} &= 4,33337 \\
 d_{2,4} &= 4,33525 \\
 d_{1,3} &= 6,85084 \\
 d_{3,5} &= 6,98268 \\
 d_{2,3} &= 8,18558 \\
 d_{1,4} &= 8,41278 \\
 d_{2,5} &= 9,66091 \\
 d_{1,5} &= 10,1635 \\
 d_{3,4} &= 10,3664 \\
 d_{4,5} &= 11,9020
 \end{aligned}$$

Next, a complete graph  $\mathbf{K}_5$  of mutual connections is constructed where the nodes are the objects  $j$ , and edges are the distances  $d_{m,n}$ .

*Taxonomical classification of hematopoietic CD34+ cell subsets...*



The third phase is the search of the disconnected graphs through successive elimination of edges which are of the largest distances.



we have first object 5, in most far distance from others

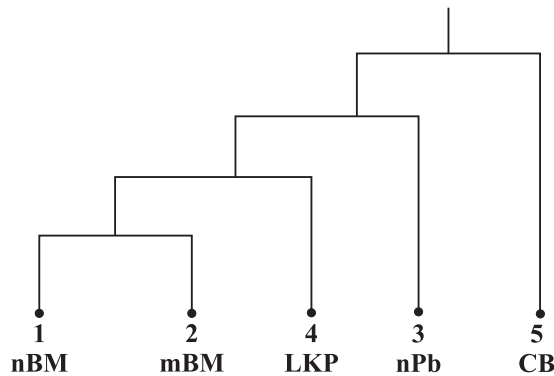
object 3, next in distance

objects 1 and 2 are closest

The resulting incoherent graphs are groups of mutually disjoint subsets containing single linkage. This may be presented as a tree with clustered objects.

Resulted classification for TABLE 1:

“Frequency of *early cell subsets* and *cellular adhesion molecules expression* within the CD34+ cell population derived from five sources of **HPC** (% of total CD34+ cells)”



**nBM** = normal bone marrow; **mBM** = mobilized bone marrow; **nPB** = normal peripheral blood; **LKP** = leukapheresis product; **CB** = cord blood

Figure above demonstrates that **LKP** is closest to **nBM** and **mBM** – these theoretical results confirm our expectations in implementing these three sources of **HPC** into transplantation.

Results for TABLE 2.

Matrix of distances  $d_{m,n}$

j →	1	2	3	4	5
↓					
1	0	3,66282	9,25248	8,67528	9,46347
2	3,66282	0	7,09278	5,89071	8,69142
3	9,25248	7,09278	0	5,76077	5,67728
4	8,67528	5,89071	5,76077	0	9,49803
5	9,46347	8,69142	5,67728	9,49803	0

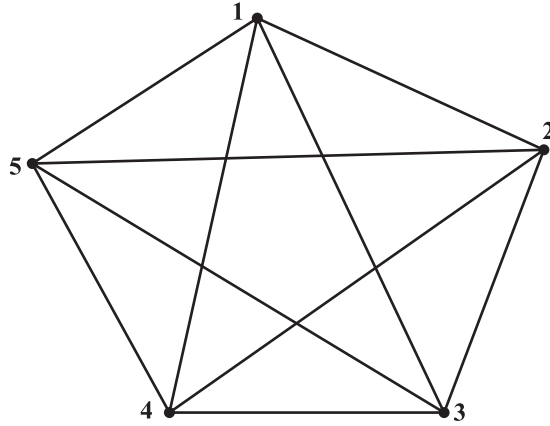
*Taxonomical classification of hematopoietic CD34+ cell subsets...*

This gives us a set of 10 distances  $\frac{j!}{r!(j-r)!} = 10$ , where  $j = 5$ ,  $r = 2$  (pair).

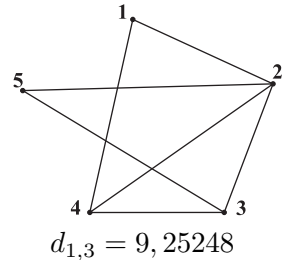
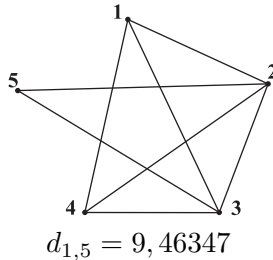
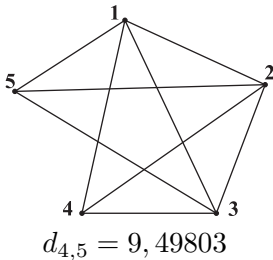
In arithmetic order

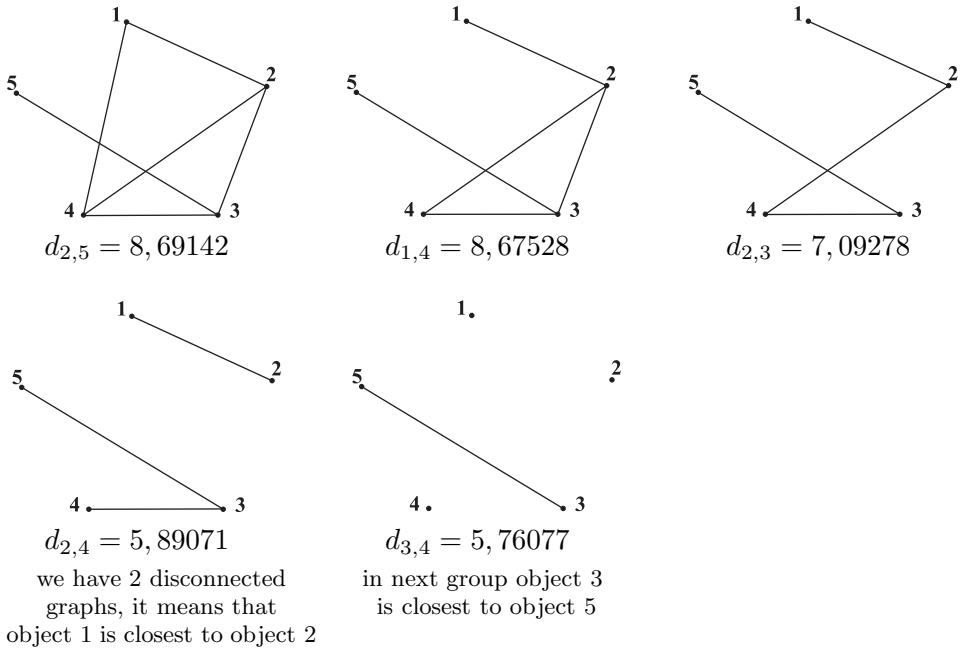
- $d_{1,2} = 3,66282$
- $d_{3,5} = 5,67728$
- $d_{3,4} = 5,76077$
- $d_{2,4} = 5,89071$
- $d_{2,3} = 7,09278$
- $d_{1,4} = 8,67528$
- $d_{2,5} = 8,69142$
- $d_{1,3} = 9,25248$
- $d_{1,5} = 9,46347$
- $d_{4,5} = 9,49803$

Next, a complete graph  $\mathbf{K}_5$  of mutual connections is constructed where the nodes are the objects  $j$ , and edges are the distances  $d_{m,n}$ .



The third phase is the search of the disconnected graphs through successive elimination of edges which are of largest distances.

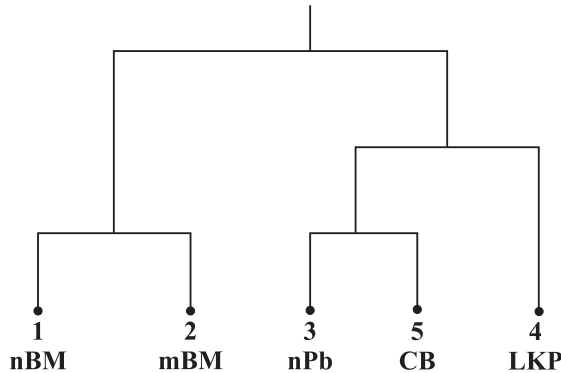




The resulting incoherent graphs are groups of mutually disjoint subsets containing single linkage. This we may be presented as a tree with clustered objects.

Resulted classification for TABLE 2:

“Antigen density of *early marker* and *cellular adhesion molecules* on CD34+ cells derived from five sources of **HPC** (Nb. of ABC x 10<sup>3</sup> molecules / cell)”



**nBM** = normal bone marrow; **mBM** = mobilized bone marrow; **nPB** = normal peripheral blood; **LKP** = leukapheresis product; **CB** = cord blood

Figure above demonstrates that **LKP** is the furthest from **nBM** and **mBM**, but it is at the same time in the group of circulated cells. The examination of features of this group enables to differentiate circulated cells vs. cells to stay in **BM**.

Results calculated with respect to standard deviation are identical as shown above.

#### R E F E R E N C E S

- [1] Matuszewski R., Trybulec A., *An algorithm for clustering in metric spaces*, Papers of the Warsaw University in Bialystok, Vol. 5, pp. 117–123, 1977.
- [2] Giero M., Matuszewski R., *Lower Tolerance. Preliminaries to Wroclaw Taxonomy*, Formalized Mathematics, Vol. 9, Nr. 3, pp. 597–603, 2001.
- [3] Sovalat H., Racadot E., Ojeda M., et al., *CD34+ cells and CD34+ CD38 – subset from mobilized blood show different patterns of adhesion molecules compared to those from steady-state blood, bone marrow and cord blood*. Journal of Hematotherapy and Stem Cell Research, 2003, Vol. 12:5, pp. 473–489.