

## DivCalc: A Utility for Diversity Analysis and Compound Sampling

Rajeev Gangal\*

SciNova Informatics, 161 Madhumanjiri Apartments, Lane 4, Dahanukar Colony, Kothrud, Pune, Maharashtra, India 411038.

\* To whom correspondence should be addressed; e-mail: [rajeev.gangal@scinovaindia.com](mailto:rajeev.gangal@scinovaindia.com)

*Received: 1 July 2002; in revised form: 29 July 2002 / Accepted: 7 August 2002 / Published: 31 August 2002*

---

**Abstract:** Diversity, in the form of genetic diversity, chemical diversity etc, is a very important concept in several areas of scientific research, and calculation of diversity is one of the most important considerations in pre-clinical drug discovery research and, in particular, in design of diverse chemical libraries for combinatorial chemistry and compound selection for High Throughput Screening (HTS). DivCalc is a Windows<sup>TM</sup> based software that implements a previously published method of diversity calculation [1]. This facilitates sampling of a given data matrix to obtain the most diverse compounds that span the entire descriptor space.

**Keywords:** Diversity, dissimilarity, software, combinatorial chemistry, HTS, DISSIM

---

### Introduction

Pharmaceutical companies routinely screen a very high number of compounds to identify ones showing biological activity (“hits”) and thus, prospective drug candidates [2-5]. One way to increase the number of compounds in corporate libraries is to use combinatorial chemistry to synthesize a large number of molecules. Diverse corporate compound libraries are thus formed by the combination of basic building blocks like carboxylic acids, amides, amines etc. The diversity in a chemical library can thus be achieved by using diverse reagents for synthesis [2, 3]. Another approach followed by many

companies is the purchase of chemicals that increase the diversity of their corporate compound collections. A HTS library can easily contain a very high number of compounds ranging between  $10^5$  and  $10^9$  [5]. The challenge is to obtain a representative but diverse sample of this collection.

Clearly, diverse compound selection is an important step in the application of combinatorial chemistry and HTS in drug discovery. The idea behind diversity based sampling is to get a diverse and representative subset of compounds. [1,5]

There are several proprietary and commercial software products and approaches for diversity analysis e.g. products like *DiverseSolutions* and *Selector* from Tripos and *Diversity Explorer* from Accelrys can select a maximally diverse set of compounds for synthesis and aid in deciding if a purchased library complements an existing corporate collection. However there is a dearth of desktop utilities with similar functionality that quickly enable assessment or sampling of diversity. *DivCalc* is an attempt to bridge this gap. It implements an algorithm (DISSIM) published earlier by Flower [1] that searches for maximally dissimilar compounds, given an input data matrix of descriptors for all the chemical compounds [1].

## Features And Implementation

DivCalc does not attempt to calculate descriptors and this paper does not address the issues behind descriptor calculation and selection. There are many commercial and free software packages available to calculate a number of molecular descriptors. Chemical Diversity Analysis literature is replete with papers that deal with issues of calculation and selection of one-, two- and three-dimensional descriptors. One excellent reference is the *Handbook of Molecular Descriptors* [6]. In general one- and two-dimensional descriptors are used for diversity analysis and three-dimensional descriptors are used for protein target specific QSAR. One such software that calculates molecular descriptors is Dragon v.2.1, available for download at <http://www.disat.unimib.it/chm/>.

The original implementation of the DivCalc algorithm was for SGI workstations with the source code written in FORTRAN77. The ubiquity of desktops in the scientific world now warrants a re-implementation for desktop users. The current implementation called *DivCalc* runs on 32 bit Windows<sup>TM</sup> operating systems. Due to memory constraints however the current version is limited to fewer than 25,000 data points. Thus the current version is useful for diversity analysis for reagent selection. The datapoint limitation is an issue with the language used for implementation and efforts are underway to use memory efficient techniques to allow for much greater number of compounds and descriptors. The input to the program is a space delimited data file with no row and column headers. The input data is shown in an Excel<sup>TM</sup> like grid (Figure 1). Transforms like Log, Transpose and Data Normalization have been provided (Figure 2). Data Normalization is a method for scaling data by using the mean and standard deviation. Sometimes the descriptors calculated have very different ranges in which they can occur. Some descriptor values can lie between 0 and 1 and others can be large real numbers. Thus the Euclidean distance can be unnecessarily biased by large values. To avoid this, Data normalization is selected by default in DivCalc and can be deselected if not needed.

Figure 1. Main interface screenshot of DivCalc

DivCalc  
File Data Analysis Diversity Analysis Help

Status: File Input Complete

	Col 1	Col 2	Col 3	Col 4	Col 5
Row 1	0.6037	0.6205	0.0612	0.4920	0.3
Row 2	0.9823	0.6404	0.2726	0.8310	0.6
Row 3	0.3988	0.7552	0.3544	0.9495	0.2
Row 4	0.1975	0.3465	0.0170	0.4504	0.9
Row 5	0.0075	0.5252	0.8349	0.3903	0.5
Row 6	0.4905	0.5423	0.9377	0.2305	0.7
Row 7	0.8336	0.5858	0.7469	0.2517	0.5
Row 8	0.1526	0.4918	0.9652	0.2120	0.5
Row 9	0.7162	0.4045	0.7230	0.5304	0.5
Row 10	0.4901	0.9137	0.2907	0.7779	0.2
Row 11	0.4531	0.7024	0.2602	0.7549	0.0
Row 12	0.9095	0.2355	0.6047	0.5390	0.6
Row 13	0.4434	0.5205	0.5207	0.5000	0.4
Row 14	0.6727	0.4721	0.4601	0.2437	0.4
Row 15	0.4575	0.9112	0.2576	0.0249	0.0
Row 16	0.7197	0.7299	0.7604	0.3132	0.7
Row 17	0.5313	0.2544	0.8658	0.2905	0.9
Row 18	0.6250	0.3055	0.3733	0.0020	0.6
Row 19	0.9862	0.1995	0.9976	0.2624	0.9
Row 20	0.6904	0.5899	0.8695	0.1624	0.7
Row 21	0.5735	0.2313	0.6637	0.5932	0.2
Row 22	0.1012	0.1472	0.3815	0.1553	0.1
Row 23	0.7434	0.8617	0.0599	0.4281	0.0
Row 24	0.3992	0.0184	0.6015	0.6850	0.8
Row 25	0.0495	0.4458	0.7126	0.7531	0.7

Figure 2. Data Analysis Menu screenshot: Transpose

DivCalc  
File Data Analysis Diversity Analysis Help

Transformations: Transpose  
Normalize  
Log Transform

	Col 1	Col 2	Col 3	Col 4	Col 5
Row 1	0.6037	0.9823	0.3988	0.1975	0.0
Row 2	0.6205	0.6404	0.7552	0.3465	0.5
Row 3	0.0612	0.2726	0.3544	0.017	0.0
Row 4	0.4920	0.8310	0.9495	0.4504	0.3
Row 5	0.3627	0.6671	0.2924	0.9067	0.5
Row 6	0.4951	0.4952	0.4699	0.3357	0.1
Row 7	0.2604	0.8999	0.1239	0.5928	0.6
Row 8	0.7895	0.9047	0.8666	0.8762	0.1
Row 9	0.2917	0.257	0.2333	0.8402	0.4
Row 10	0.3247	0.5295	0.4537	0.0716	0
Row 11	0.682	0.6721	0.0163	0.8095	0.7
Row 12	0.514	0.0389	0.6016	0.3923	0.5
Row 13	0.9881	0.0314	0.6919	0.4742	0.8
Row 14	0.1119	0.8178	0.3854	0.5773	0.3
Row 15	0.2322	0.5392	0.4016	0.2038	0.8
Row 16	0.9056	0.5151	0.5938	0.0713	0.0
Row 17	0.956	0.5207	0.2028	0.005	0.7
Row 18	0.7731	0.1451	0.5957	0.2835	0.4
Row 19	0.3767	0.5251	0.9265	0.566	0.5
Row 20	0.2212	0.5934	0.3975	0.7944	0.5
Row 21	0.7622	0.3933	0.1187	0.7113	0.2
Row 22	0.2934	0.8637	0.4781	0.1654	0.2
Row 23	0.6022	0.3008	0.5952	0.6548	0.4
Row 24	0.3747	0.9681	0.1057	0.7331	0
Row 25	0.7472	0.743	0.7252	0.0994	0.4

The main functionality in the menu is of course to select the most diverse compounds from the dataset. The user can set a predefined limit on the number or percentage of compounds to be displayed in the output. By default, all rows/compounds in the input are ranked by their diversity and shown in the output (Figure 3).

**Figure 3.** Diversity Analysis screenshot

The screenshot shows the DivCalc software interface. The main window displays the results of a Diversity Analysis. The window title is "DivCalc" and the menu bar includes "File", "Data Analysis", "Diversity Analysis", and "Help". The status bar indicates "Status: Diversity Analysis Results". The main display area contains two tables of data, each with 5 columns labeled "Col 1" through "Col 5". The left table shows rows 1 through 25, and the right table shows rows 67 through 3. The data values are numerical, representing the results of the diversity analysis.

Row	Col 1	Col 2	Col 3	Col 4	Col 5
Row 1	0.6037	0.6205	0.8612	0.452	0.3
Row 2	0.5823	0.6404	0.2726	0.831	0.6
Row 3	0.3938	0.7552	0.3544	0.9495	0.2
Row 4	0.1976	0.3465	0.017	0.4504	0.5
Row 5	0.0079	0.5252	0.8349	0.3903	0.5
Row 6	0.4805	0.5423	0.9377	0.2305	0.7
Row 7	0.6336	0.5858	0.7469	0.2517	0.5
Row 8	0.1526	0.4918	0.9652	0.212	0.5
Row 9	0.7162	0.4045	0.723	0.5304	0.5
Row 10	0.4901	0.9137	0.2907	0.7779	0.1
Row 11	0.4531	0.7024	0.2502	0.7549	0.0
Row 12	0.9035	0.2395	0.6047	0.939	0.6
Row 13	0.4434	0.5205	0.5207	0.593	0.4
Row 14	0.6727	0.4721	0.4601	0.2437	0.4
Row 15	0.4575	0.9112	0.2576	0.0249	0.0
Row 16	0.7197	0.7299	0.7504	0.3132	0.7
Row 17	0.5313	0.2544	0.8558	0.2905	0.9
Row 18	0.625	0.3395	0.3733	0.002	0.6
Row 19	0.9662	0.1895	0.5576	0.2624	0.9
Row 20	0.6804	0.5899	0.8585	0.1824	0.7
Row 21	0.5735	0.2313	0.6537	0.9332	0.2
Row 22	0.1012	0.1472	0.3815	0.1553	0.1
Row 23	0.7434	0.8617	0.0589	0.4281	0.1
Row 24	0.3532	0.0184	0.6015	0.685	0.8
Row 25	0.0975	0.4459	0.7126	0.2531	0.7
Row 67	0.2151	0.9218	0.078	0.8977	0.7
Row 6	0.4809	0.5423	0.9377	0.2305	0.7
Row 13	0.4434	0.5205	0.5207	0.593	0.4
Row 82	0.6408	0.7851	0.1595	0.3638	0.0
Row 66	0.8843	0.9497	0.3234	0.8677	0.6
Row 75	0.5018	0.9458	0.9874	0.3267	0.6
Row 25	0.6389	0.2008	0.8849	0.6586	0.5
Row 95	0.0049	0.113	0.9013	0.0023	0.3
Row 35	0.5683	0.1879	0.9452	0.3717	0.8
Row 61	0.4823	0.7416	0.3127	0.4352	0.4
Row 4	0.1576	0.3465	0.017	0.4504	0.9
Row 10	0.4901	0.9137	0.2907	0.7779	0.1
Row 78	0.7115	0.7824	0.8751	0.086	0.2
Row 99	0.4615	0.987	0.3369	0.1555	0.6
Row 69	0.5217	0.7675	0.3815	0.095	0.2
Row 63	0.5842	0.2713	0.3345	0.9158	0.1
Row 30	0.1836	0.9557	0.9573	0.4498	0.6
Row 8	0.1526	0.4918	0.9652	0.212	0.5
Row 52	0.5119	0.1894	0.1004	0.4581	0.8
Row 49	0.1162	0.419	0.9267	0.506	0.9
Row 35	0.0174	0.2581	0.0821	0.1197	0.5
Row 22	0.1012	0.1472	0.3815	0.1553	0.1
Row 18	0.625	0.3355	0.3733	0.002	0.6
Row 5	0.0079	0.5252	0.8349	0.3903	0.5
Row 3	0.3908	0.7552	0.3544	0.9495	0.2

## Algorithm

The algorithm selects  $k$  most diverse compounds from a given set of  $m$  compounds where  $k \leq m$ . All calculations are performed using Euclidean distance as a measure of dissimilarity.

1. The centroid of the input data is calculated.
2. The compound most distant from the centroid is the first selected compound. Thus the compound with the maximum Euclidean distance from the centroid is the first selected compound.
3. The compound most distant from the 1<sup>st</sup> compound is selected next.
4. Hereafter the algorithm increases in complexity. At each iteration the procedure is to find the compound from the list of unselected compounds whose minimum distance to selected compounds is a maximum amongst all unselected compounds.
5. Step 4 is repeated till the required number of compounds are selected.

A visual reproduction of the above steps can be seen in the paper describing the original implementation [1].

## Discussion

*DivCalc* is a desktop program to calculate diversity for a set of compounds and to select a subset of maximally diverse compounds. The algorithm has been implemented with the desktop user in mind working in a Windows<sup>TM</sup> environment. Work is in progress to enhance the number of data points that can be handled and provision of alternative distance formulae like Manhattan distance, correlation etc.

When *DivCalc* is used to find a small number of compounds, the algorithm identifies a set that is broadly spread over the whole descriptor space. However as the size of the subset increases, the dissimilarity of newly selected compounds to ones already selected, decreases rapidly[1]. So, another feature being currently worked upon is a default calculation of the ideal sample size of the subset that should be drawn. Copies of *DivCalc* can be obtained by sending a request to [rajeev.gangal@scinovaindia.com](mailto:rajeev.gangal@scinovaindia.com)

## References

1. Flower, D. DISSIM: A program for the analysis of chemical diversity. *J. Mol. Graphics Mod.* **1998**, *16*, 239-253.
2. Van Drie, J. H.; Lajiness M. S. Approaches to virtual library design. *DDT*, **1998**, *3*, 274-283.
3. Linusson, A.; Gottfries, J.; Lindgren, F.; Wold S. Statistical Molecular Design of Building Blocks for Combinatorial Chemistry. *J. Med. Chem.*, **2000**, *43*, 1320-1328.
4. Leach, A. R.; Hann, M. M. The *in silico* world of virtual libraries. *DDT*, **2000**, *5*, 326-336.
5. Gorse, D.; Rees, A.; Kaczorek, M; Lahana, R. Molecular Diversity and its analysis. *DDT*, **1999**, *4*, 257-385.
6. Todeschini, R; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim (Germany), 2000; vol. 11, pp. 667.

*Sample availability*: Not applicable

© 2002 by MDPI (<http://www.mdpi.org>). Reproduction is permitted for noncommercial purposes.