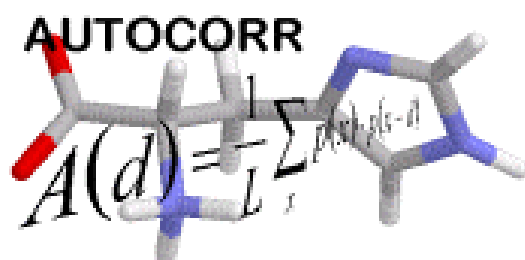


AUTOCORR

Calculation Autocorrelation Coefficients of Molecules Encoding Physicochemical Atom Properties

Version 1.1

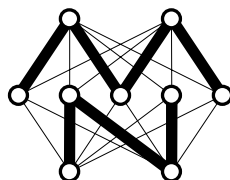
Program Description



Christof H. Schwab and Johann Gasteiger

Molecular Networks GmbH Computerchemie
October 2002

<http://www.mol-net.de>



Molecular Networks GmbH
Computerchemie
Nägelsbachstr. 25
91052 Erlangen
Germany

Phone: +49-(0)9131-815668
Fax: +49-(0)9131-815669

Email: info@mol-net.de
WWW: www.mol-net.de

This document is copyright © 2001 by Molecular Networks GmbH Computerchemie. All rights reserved. Except as permitted under the terms of the Software Licensing Agreement of Molecular Networks GmbH Computerchemie, no part of this publication may be reproduced or distributed in any form or by any means or stored in a database retrieval system without the prior written permission of Molecular Networks GmbH Computerchemie.

The software described in this document is furnished under a license and may be used and copied only in accordance with the terms of such license.

CORINA and **PETRA** are registered trademark in the Federal Republic of Germany. Other product names and company names may be trademarks or registered trademarks of their respective owners, in the Federal Republic of Germany and other countries. All rights reserved.

Contents

1	Program Installation	1
1.1	New Installation	1
1.2	Program Updates	2
2	Problems and Help!	3
3	Getting Started	4
4	Program Use	5
5.1	Synopsis	5
5.2	Options	5
5	Tutorial: How to Calculate Autocorrelation Coefficients of Molecules Encoding Physicochemical Atom Properties	7
5.1	Prepare your structure file	7
5.2	Run CORINA	8
5.3	Run PETRA	8
5.4	Run AUTOCORR	9
6	Acknowledgments	12
13	References	13
14	Report Form	14

1 Program Installation

Since version 1.1 AUTOCORR is distributed on a CD-ROM, which contains the executable file(s) of AUTOCORR, this program description in PDF format, and an example file of structure and physicochemical information (see chapter 3 on page 4).

The CD-ROM contains an ISO9660 file system and, thus, is readable by all common UNIX systems as well as by Microsoft Windows (win32) based platforms. The following directories and files are common for all hardware platforms.

name of directory	description	name of file(s)
examples	example files for structure data (ASCII)	<i>example.ctx</i>
manual	this program description in PDF format	<i>autocorr11manual.pdf</i>

Please copy the example file *example.ctx* into your home directory. The program description *autocorr11manual.pdf* can be viewed and printed with a PDF document viewer, e.g. Adobe Acrobat Reader (<http://www.adobe.com/acrobat>).

1.1 New Installation

1.1.1 Sun SPARC Stations

The directory *sun* on the CD-ROM contains the Sun Solaris executable file *autocorr.sun* and a configuration file for AUTOCORR *autocorr.cfg* (ASCII).

- 1) Create a subdirectory, e.g., *autocorr* (for system administrators when installing software locally, e.g. */usr/local/bin/autocorr*).
- 2) Copy the file *autocorr.sun* from the CD-ROM to the subdirectory *autocorr* and rename the file *autocorr.sun* to *autocorr*. **Please note:** *autocorr.sun* is a binary file.
- 3) Copy the file *autocorr.cfg* from the CD-ROM to your working directory from which AUTOCORR will be called. **Please note:** *autocorr.cfg* is an ASCII file.
- 4) Add the *autocorr* subdirectory name to the environment variable *PATH* in your *.login* or *.cshrc* files.

1.1.2 Silicon Graphics

The directory *sgi* on the CD-ROM contains the Silicon Graphics executable file *autocorr.sgi* and a configuration file for AUTOCORR *autocorr.cfg* (ASCII).

- 1) Create a subdirectory, e.g., *autocorr* (for system administrators when installing software locally, e.g. */usr/local/bin/autocorr*).
- 2) Copy the file *autocorr.sgi* from the CD-ROM to the subdirectory *autocorr* and rename the file *autocorr.sgi* to *autocorr*. **Please note:** *autocorr.sgi* is a binary file.
- 3) Copy the file *autocorr.cfg* from the CD-ROM to your working directory from which AUTOCORR will be called. **Please note:** *autocorr.cfg* is an ASCII file.
- 4) Add the *autocorr* subdirectory name to the environment variable *PATH* in your *.login* or *.cshrc* files.

1.1.3 IBM Compatible Systems - Linux

The directory *linux* on the CD-ROM contains the Linux executable file *autocorr.lnx* and a configuration file for AUTOCORR *autocorr.cfg* (ASCII).

- 1) Create a subdirectory, e.g., *autocorr* (for system administrators when installing software locally, e.g. */usr/local/bin/autocorr*).
- 2) Copy the file *autocorr.lnx* from the CD-ROM to the subdirectory *autocorr* and rename the file *autocorr.lnx* to *autocorr*. **Please note:** *autocorr.lnx* is a binary file.
- 3) Copy the file *autocorr.cfg* from the CD-ROM to your working directory from which AUTOCORR will be called. **Please note:** *autocorr.cfg* is an ASCII file.
- 4) Add the *autocorr* subdirectory name to the environment variable *PATH* in your *.login* or *.cshrc* files (*.profile* or *.bashrc*).

1.2 Program Updates

- 1) Before installing the new version, please copy the old executable and configuration files to a new directory, e.g. *autocorrxxx* (*xxx* = *old-version-number*, e.g., *autocorr10*).
- 2) According to the hardware platform install the new version following the instructions given in section 1.1 on page 1.

2 Problems and Help!

If you have any difficulties with the installation of AUTOCORR or if any problems occur while running AUTOCORR, please send all your inquiries to the following address:

Molecular Networks GmbH Computerchemie
Nägelsbachstr. 25
91052 Erlangen
Germany

or contact us by email support@mol-net.de,

or by Fax +49-(0)9131-815669.

Please include your input file, the output file, and the AUTOCORR trace file *autocorr.trc* generated by AUTOCORR on an MS/DOS diskette (3½") or send it to us by email. These files will help us to analyze the problem; if your system displays any error messages, please add them to your report. Thank you!

You can also use the report form at the end of this manual.

3 Getting Started

The example file *example.ctx* submitted with the distribution contains the structure information of three molecules in CTX file format, [1] the default file format of AUTOCORR. Copy this example file into your working directory and type the following command:

```
autocorr -s -H -2 -w3d -pautocorr.cfg example.ctx out.ctx
```

the following output appears on the screen:

```
autocorr <serial number - compilation date - user - date - time - host>
```

```
INFORMATION autocorr(): program settings:
```

```
    excluding hydrogen atoms
    excluding 2D autocorrelation coefficients
    weighting 3D autocorrelation coefficients
    using property file autocorr.cfg
    using property/ies:
A_QSIG A_QPI A_QTOT A_POLARIZABILITY
```

```
*** RECORD no.:  1 read *****
    Ident 3
    Name metoprolol
    Elapsed time: 0 ms
```

```
*** RECORD no.:  2 read *****
    Ident 3
    Name nordiazepam
    Elapsed time: 0 ms
```

```
.
.
.
```

```
*** RECORD no.: 17 read *****
    Ident 1
    Name sulfasalazine
    Elapsed time: 0 ms
```

```
*** RECORD no.: 18 read *****
    Ident 1
    Name olsalazine
    Elapsed time: 0 ms
```

```
18 record(s) read, 18 converted
Totally elapsed time: 0 s
```

4 Program Use

4.1 Synopsis

The general synopsis for using AUTOCORR is:

```
autocorr [ -option[value] -option[value] ... ] [ infile ] [ outfile ]
```

Infile and outfile are the input and output file names. If no file names are given, the program reads from standard input and writes to standard output. If only one file name is given, this file will be read as input file and the output will be written to standard output. Minimum trace output is by default written to the file *autocorr.trc*.

4.2 Options

The command line options follow the rules of the UNIX command syntax standard.

- a append the output to the output file (default: overwrite)
- s write the trace information to the standard output channel
(default: autocorr.trc)
- wtsv extract the autocorrelation coefficients in tsv format (tab
separated values for spread sheet or KMAP input [2], default:
ctx)
- n<value> apply a maximum topological distance of <value> for the
calculation of topological autocorrelation coefficients (<value>
of integer type, default: 10 topological distances)
- H exclude the hydrogen atoms from the calculation of both
topological and spatial autocorrelation coefficients
- l<value> set the lower 3D distance border *l* for the calculation of spatial
autocorrelation coefficients to <value> Å (<value> of float
type, default: 1 Å)
- u<value> set the upper 3D distance border *u* for the calculation of spatial
autocorrelation coefficients to <value> Å (<value> of float
type, default: 13 Å)

- I<value> set the number of equal 3D distance intervals I for the calculation of spatial autocorrelation coefficients to <value> (<value> of integer type, default: 12 intervals)
- w2d weight the 2D (topological) autocorrelation coefficients by the total number of distances found in the interval under consideration
- w3d weight the 3D (spatial) autocorrelation coefficients by the total number of distances found in the interval under consideration
- 2 skip the calculation of the 2D (topological) autocorrelation coefficients
- 3 skip the calculation of the 3D (spatial) autocorrelation coefficients
- f fuzzify the 3D (spatial) autocorrelation
- W<value> set the width W of the fuzzified overlap to <value> (interval width) (<value> of integer type, only with option -f)
- p<value> use the atomic properties specified in the property file <value> for the calculation of autocorrelation coefficients (<value> of character type, default: A_QSIG, A_QPI, A_QTOT, A_ENSIG, A_ENPI, A_ENLP, A_POLARIZABILITY, A_A1)
- help print the online help to the standard output channel
- ver Print program version

5 Tutorial: How to Calculate Autocorrelation Coefficients of Molecules Encoding Physicochemical Atom Properties

This short tutorial guides through the process of calculating a set of molecular descriptors based on autocorrelation functions which encode physicochemical properties. The procedure described below can of course be adapted to the needs of any study. A detailed description on the usage of the program systems mentioned in this tutorial can be found in the corresponding program manuals.

The goal of this tutorial is to show how to calculate topological (2D) and spatial (3D) autocorrelation coefficients encoding physicochemical atom properties with the software packages provided by Molecular Networks. Autocorrelation functions can be used for structure representation in order to obtain molecular descriptors, e.g. for the classification of chemical compound libraries or for the prediction of any molecular property of interest.

An overview on the basic principles of autocorrelation and its application in chemistry and chemoinformatics can be obtained in [3].

Molecular Networks also offers the software package SURFACE for deriving autocorrelation coefficients encoding molecular surface properties (e.g., the molecular electrostatic potential, MEP) [4]. This is described in the tutorial "**How to Calculate Autocorrelation Coefficients of Molecules Encoding Molecular Surface Properties**" and in the program manual of SURFACE.

Depending on the level of molecular information which should be incorporated into a study, two or three different program packages are required to obtain a set of autocorrelation coefficients for a molecular structure.

For the derivation of topological (2D) autocorrelation coefficients:

- PETRA to calculate physicochemical properties of atoms [5]
- AUTOCORR to derive the autocorrelation coefficients

To obtain spatial (3D) autocorrelation coefficients, three-dimensional information of the molecules under consideration is required and, therefore, the 3D Structure Generator CORINA has to be run first:

- CORINA to generate 3D structures [6]
- PETRA to calculate physicochemical properties of atoms
- AUTOCORR to derive the autocorrelation coefficients

5.1 Prepare your structure file

Prepare a multi record SDF file (either 2D or 3D) or a multi record file containing the SMILES strings of your molecules under investigation.

The SDF file should contain the names of the compounds in the first line of the header block or in an appropriate non-structural data item field (after the `M END` keyword), e.g.

```
> <COMPOUND.NAME>  
Cyclohexanol
```

Structure files containing SMILES codes should contain only one compound (SMILES string) per line. The first string in a line is considered to be the SMILES code and all following strings in the same line are interpreted as compound names, e.g.

```
C1CCCCC1(O) cyclohexanol
```

For both cases, SDFfile or SMILES input, please ensure that the stereochemical information is specified correctly. Note that stereochemistry can highly influence a certain property exhibited by a chemical compound.

5.2 Run CORINA

To generate 3D structures start CORINA with the following command line options:

For SDFfiles with the compound name in the header block:

```
corina -i t=sdf -o t=sdf -d wh,r2d,rs infile.sdf outfile3D.sdf
```

For SDFfiles with the compound name in a non-structural data item field:

```
corina -i t=sdf,sdfi2n=COMPOUND.NAME -o t=sdf -d wh,r2d,rs infile.sdf  
outfile3D.sdf
```

For SMILES input:

```
corina -i t=smiles,smilesname -o t=sdf -d wh,r2d,rs infile.smi  
outfile3D.sdf
```

The options starting with `-i` and `-o`, respectively, force CORINA to read in and to output the specified file formats and to copy the compound names into the correct data fields in the output file.

The driver options starting with `-d` force CORINA to output the hydrogen atoms, which are given implicitly in the input file and are added during the 3D generation process (option `wh`). Furthermore, structures are removed from the output file for which no 3D model can be generated (option `r2d`), and small fragments are removed from the output file, e.g. counter ions in salts (option `rs`).

Information on the CORINA run can be found in the trace file `corina.trc`, which is automatically written into your working directory.

CORINA offers many more options to influence the 3D generation process. All available options are described in detail in the program manual of CORINA.

5.3 Run PETRA

To calculate the atom properties start PETRA with the following command line options:

```
petra -i t=sdf infile3D.sdf -o t=ctx outfile3D_petra.ctx -l petra.trc  
-c petra.cfg
```

Again, the options starting with `-i` and `-o`, respectively, force PETRA to read in and to output the specified file formats. The output file format is CTX (Gasteiger Clear Text file format, ASCII format), which is briefly described in the program manuals of CORINA and PETRA. This file format is required to be able to continue in the next step with the

AUTOCORR program package.

Information on the PETRA run can be found in the trace file `petra.trc`, which is automatically written into your working directory forced by the option `-l`.

The option `-c` forces PETRA to use the driver options given in the PETRA configuration file `petra.cfg` (ASCII). An example configuration file is distributed with PETRA (`petra.cfg`). You can use this configuration file for the task described here or adapt it to your needs. Please note that at least the statement `A_XYZ=write` has to be specified in a configuration file. Otherwise, the 3D coordinates calculated in the previous CORINA run would get lost.

The format and options of a PETRA configuration file are described in the program manual of PETRA in section 5.3.3.

The configuration file `petra.cfg`, which is distributed with PETRA, forces the program to calculate the following atom properties (see Table 1).

Table 1: Atomic properties calculated by PETRA using the configuration file `petra.cfg`

property	property name	unit
effective atom polarizability	A_POLARIZABILITY	Å ³
σ-charge	A_QSIG	electron units
σ-electronegativity	A_ENSIG	eV
π-charge	A_QPI	electron units
total charge	A_QTOT	electron units
π-electronegativity	A_ENPI	eV
lone pair electronegativity	A_ENLP	eV

A complete list of the properties which can be calculated with PETRA is shown in the program manual of PETRA in chapter 6.

The output file `outfile3D_petra.ctx` now contains the 3D information and the atom properties which are needed to calculate the autocorrelation coefficients in the following step. In the program manual of PETRA a description of the syntax for storing the calculated atom properties in a CTX file is given.

5.4 Run AUTOCORR

To derive the autocorrelation coefficients start AUTOCORR with the following command line options:

For the calculation of topological autocorrelation coefficients only:

```
autocorr -wtsv -3 -H -w2d -pautocorr.cfg file3D_petra.ctx  
file_autocorr2D.tsv
```

For the calculation of spatial autocorrelation coefficients only:

```
autocorr -wtsv -2 -H -w3d -pautocorr.cfg file3D_petra.ctx  
file_autocorr3D.tsv
```

For the calculation of both topological and spatial autocorrelation coefficients:

```
autocorr -wtsv -H -w3d -w3d -pautocorr.cfg file3D_petra.ctx  
file_autocorr.tsv
```

The default output file format of AUTOCORR is the CTX format, which is briefly described in the program manuals of CORINA and PETRA. All command lines shown above force AUTOCORR to output a file in tsv format (tab separated value format, option `-wtsv`) containing some information in the header lines about the used atom properties, number of coefficients and preset topological or spatial distances followed by the autocorrelation coefficients for each molecule. Each line contains the coefficients for one single molecule and ends with the identification number of the molecule and the compound name (with a leading !). An example output is given below.

```
!! autocorr tsv format  
!! 3D autocorr coeffs: AC3D  
!! minimum distance for AC3D in A: 1.00  
!! maximum distance for AC3D in A: 13.00  
!! number of intervals for AC3D: 12  
!! properties used for AC3D: A_QTOT  
!! number of AC3D coeffs: 12  
AC3D_A_QTOT_1 AC3D_A_QTOT_2 AC3D_A_QTOT_3 AC3D_A_QTOT_4 AC3D_A_QTOT_5  
AC3D_A_QTOT_6 AC3D_A_QTOT_7 AC3D_A_QTOT_8 AC3D_A_QTOT_9  
AC3D_A_QTOT_10 AC3D_A_QTOT_11 AC3D_A_QTOT_12 Ident !Name  
-6.873339e-04 1.286262e-03 7.699167e-05 7.737651e-04 2.242102e-04  
2.125720e-04 9.707951e-04 -9.118148e-05 -1.345536e-04 7.053868e-04  
7.097302e-04 2.416514e-05 1 !metoprolol  
-1.801498e-03 -2.941126e-05 7.428863e-04 1.577254e-04 4.582458e-05 -  
2.340810e-04 7.980547e-04 0.000000e+00 0.000000e+00 0.000000e+00  
0.000000e+00 0.000000e+00 2 !nordiazepam  
-1.603302e-03 -6.652727e-05 6.549577e-04 1.353805e-04 3.280392e-05 -  
2.246445e-04 7.164332e-04 0.000000e+00 0.000000e+00 0.000000e+00  
0.000000e+00 0.000000e+00 3 !diazepam
```

In the example above, the spatial autocorrelation coefficients from 1 to 13 Å have been calculated for three compounds and were sampled in twelve intervals (default options of AUTOCORR). Thus, twelve coefficients are obtained. The encoded atom property is the total atom charge q_{tot} (A_QTOT) as calculated by the previous PETRA run.

For the calculation process the hydrogen atoms of each molecule were not taken into account (option `-H`) and the autocorrelation coefficients are weighted by the total number of distances occurring in the interval under consideration (option `-w2d` for topological autocorrelation and option `-w3d` for spatial autocorrelation).

Information on the AUTOCORR run can be found in the trace file `autocorr.trc`, which is automatically written into your working directory.

The output file of AUTOCORR in tsv format can easily be imported into a standard spread sheet program (e.g. Microsoft Excel) in order to merge the molecular descriptors with any property data, e.g. biological activity or physicochemical data. It can also directly be used as input file for the Kohonen neural network generator KMAP, which is also distributed by Molecular Networks. In these cases, the `ident` field in the tsv file should already contain the property under investigation.

AUTOCORR offers many more options to influence the calculation of the autocorrelation coefficients. One option is how to determine which physicochemical properties should be used for encoding. This is managed by the configuration file `autocorr.cfg`. An example for a configuration file `autocorr.cfg` is distributed with the program package AUTOCORR. If no configuration file is specified (option `-p`) AUTOCORR by default uses all properties listed in Table 1 of this tutorial, plus the atom identity (which is simply the value of 1 for each atom regardless of the atom type).

The number of topological (2D) autocorrelation coefficients which are calculated can be influenced with the command line options `-n<value>`, where `<value>` is the maximum topological distance under consideration (the maximum number of intervening bonds following the shortest path which connects two atoms).

The number of spatial (3D) autocorrelation coefficients which are calculated can be influenced with the command line options

`-l<value>`, where `<value>` is the smallest 3D distance (in [Å]) between two atoms taken into account,

`-n<value>`, where `<value>` is the largest 3D distance (in [Å]) which is used, and

`-I<value>`, where `<value>` is the number of equidistant intervals in which the occurring 3D atom distances are sampled.

6 Acknowledgments

AUTOCORR was developed in the research group of Prof. Johann Gasteiger at the Technical University of Munich and at the University of Erlangen-Nürnberg since 1996. The program development was initiated by Dr. Jens Sadowski. AUTOCORR is implemented in C programming language.

AUTOCORR is now maintained for general usage by Molecular Networks GmbH Computerchemie (by Dr. Christof H. Schwab).

7 References

- [1] Gasteiger, J. et al. CTX Keyword Reference Manual. University of Erlangen-Nürnberg: 1995, unpublished results.
- [2] (a) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, Second Edition, Wiley-VCH, Weinheim, 1999, ISBN 3-527-29779-0. (b) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Sadowski, J.; Teckentrup, A.; Wagener, M. The Use of Self-Organizing Neural Networks in Drug Design. In *3D QSAR in Drug Design - Volume 2*; Kubinyi, H.; Folkers, G.; Martin, Y.C. (Eds.) Kluwer/ESCOM, Dordrecht, NL, 1998, pp 273-299. (c) The neural networks package SONNIA Version 4.10 is available from Molecular Networks GmbH, Erlangen Germany (<http://www.mol-net.de>).
- [3] Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205-1213.
- [4] (a) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769-7775. (b) SURFACE Version 1.10 is available from Molecular Networks GmbH, Erlangen Germany (<http://www.mol-net.de>).
- [5] Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Jochum, C.; Hicks, M.G.; Sunkel, J. Eds.; Springer-Verlag: Heidelberg; 1988, pp 119-138. (b) PETRA Version 3.05 is available from Molecular Networks GmbH, Erlangen Germany (<http://www.mol-net.de>).
- [6] (a) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Model Builders. *Chemical Reviews* **1993**, *93*, 2567-2581. (b) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000-1008. (c) CORINA Version 2.64 is available from Molecular Networks GmbH, Erlangen Germany (<http://www.mol-net.de>).

8 Report Form

In the case of problems occurring during installation or running AUTOCORR, please complete the following form and send it or fax it to

Molecular Networks GmbH Computerchemie
Nägelsbachstr.25
91052 Erlangen, Germany
FAX: +49-(0)9131-815669

User:

AUTOCORR program and version number ("autocorr -ver"):

Command line to run AUTOCORR:

Error and warning messages by AUTOCORR:

System messages:

Short description:

Please include the input file, output file and trace file (*autocorr.trc*) generated by AUTOCORR on a 3½" diskette written in MS/DOS format or forward it via email to support@mol-net.de These files will help us to analyze your problems. All data will be treated confidentially.