

---

# **Bioinfo-C**

## **DebianMed Meeting Aberdeen 2014**

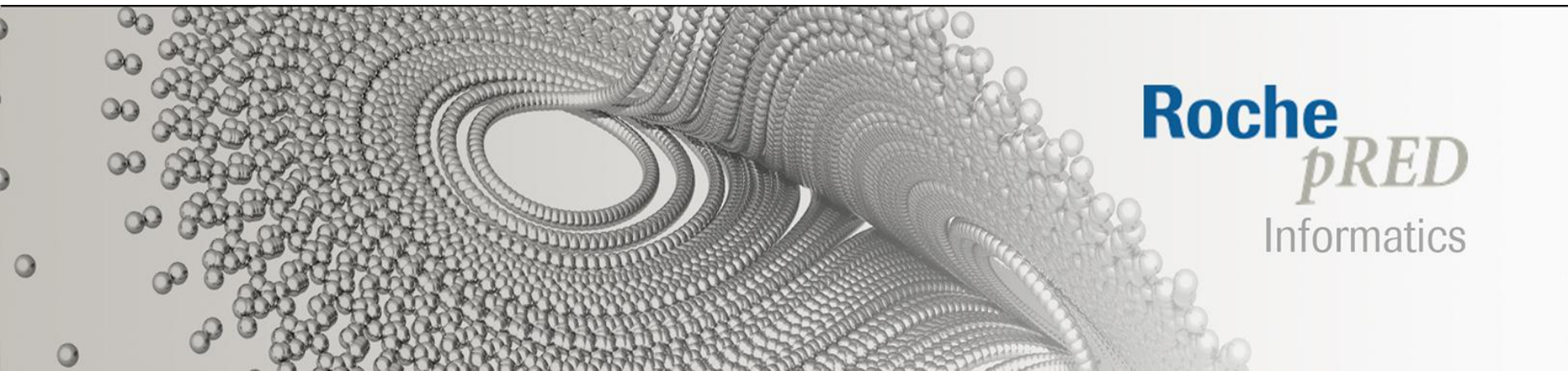
*Clemens Broger*

*Jitao David Zhang*

*Detlef Wolf*

2014-01-28

contact: [Detlef.Wolf@roche.com](mailto:Detlef.Wolf@roche.com)



# Detlef.Wolf@Roche.com

- ~1978 started programming in BASIC
- 1998 finished Diploma in Informatics @ TH Darmstadt, Germany
- 1990 – 1995 German Cancer Research Centre, Heidelberg  
(Integrated Genomic Database, Otto Ritter, → ACeDB)
- since 1996  
F. Hoffmann La Roche, Basel  
Bacterial Genomics → Toxicogenomics → Bioinformatics



**bioinfoc.ch**

## Home page

User: wolfd

[Sign out](#) | [Contact](#) | [About](#)

## Applications

[Pubmed search](#)  
[Sequence Analysis Web Interface](#)  
[GeneLookup](#)  
[CompoundLookup](#)  
(more under development)

## Demonstrations

[Webfile](#)  
[BioinfoLib](#)

## Bioinfo-C is

integrated free (LGPL) object based  
C code to efficiently develop  
bioinformatics applications.  
[More ...](#)

## Download

from [sourceforge.net](#)

## Documentation

[Bioinfoc Wiki](#)  
[BioinfoLib man pages](#)  
[BioinfoLib tutorials](#)

## Admin



# Topics

- Bioinfo-C is ...
- selected topics
  - BioinfoLib.kern
  - PubmedSearch
  - Ribios



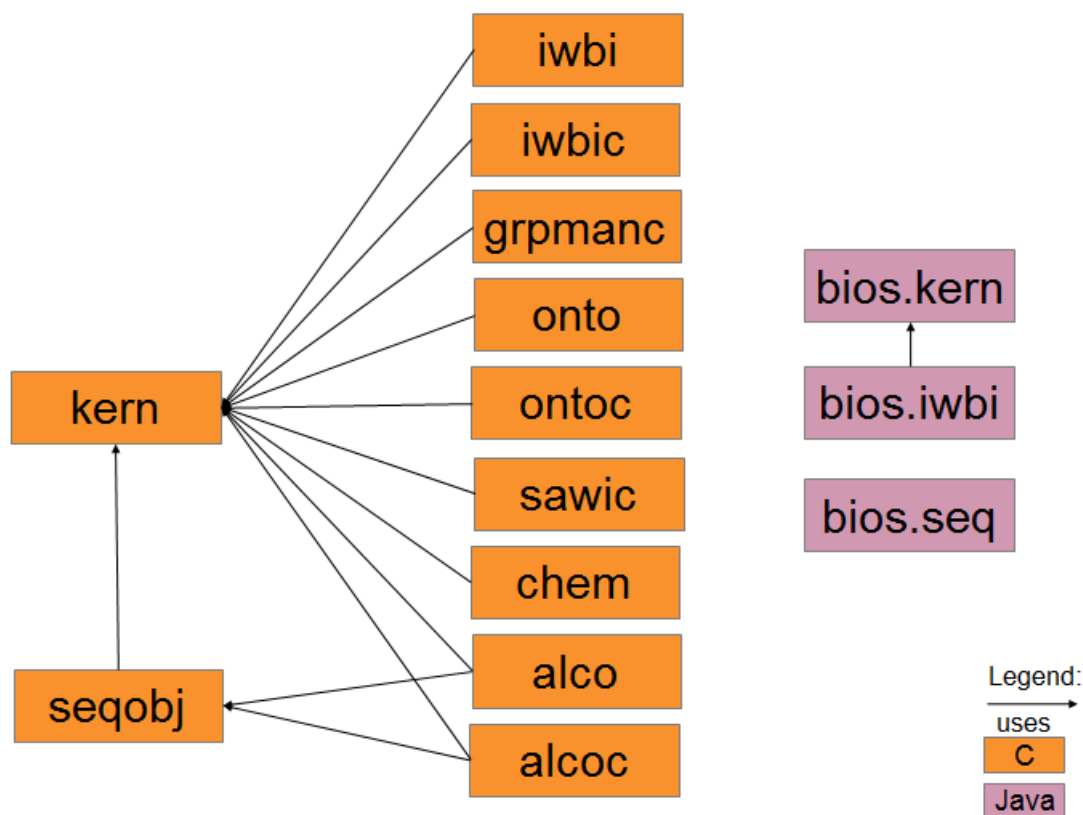
doxygen



# Bioinfo-C is ...

- Bioinfo-C is integrated free (LGPL) object based code in C, Java and R. Bioinfo-C consists of a central **library** called BioinfoLib written mainly in C but also containing some Java packages and R libraries. BioinfoLib allows to efficiently develop bioinformatics (and other) applications in C, Java and R. The Webfile framework and some **applications** are written using BioinfoLib and are also part of Bioinfo-C.
- Bioinfo-C originates from F.Hoffmann-La Roche AG, with headquarters in Basel, Switzerland.
- Many individuals helped Bioinfo-C grow to its present state, especially: Axel Klenk, Bjoern Gaiser, Martin Strahm, Martin Ebeling, Laura Badi, Isabelle Wells, Guido Steiner, Yuan Wang, Said Aktas, Marco Berrera, Roland Schmucki, David Zhang, Klaus Weymann, Liping Jin, Lukas Habegger, Abhishek Garg, James Holzwarth, Xing Yang, Jay Vaymarani, Sittichoke Saisanit, Shalin Tang, Martin Schweiger, Marc Lieber, Daniel Doran, Martin Neeb, Michael Braxenthaler, Detlef Wolf and Clemens Broger.

# BioinfoLib Subsystems



Subsystem	Functionality
<b>kern</b>	<b>basic functionality, various parsers, etc.</b>
iwbi	user identification (authentication)
iwbic	client to iwbi
grpman	client to grpman server (authorization)
onto	server that can hold ontologies as DAGs
ontoc	client to query ontologies of onto
sawic	client that connects to the sawi server
chem	functionality to handle SMILES, molfiles, etc.
alco	server that handles alignment sets (e.g. serialized Blast searches, etc.)
alcoc	client to access the alco server
seqobj	functionality to handle sequences and alignments in various formats

# BioinfoLib.kern -- 70 modules

## Selected modules:

[http://bioinfoc.ch/doxygen/bios/html/dir\\_b3dad8e81b3f37b89b2bf0a8abb2d993.html](http://bioinfoc.ch/doxygen/bios/html/dir_b3dad8e81b3f37b89b2bf0a8abb2d993.html)

Module	Purpose
affyfileHandler.c	Knows how to parse several kinds of Affymetrix files. Module prefixes cfo_, lfo_, ifo_, dfo_, efo_.
algotil.c	Makes local or global sequence alignments. Uses code from EMBOSS. Module prefix algotil_.
<b>array.c</b>	Module for handling dynamic arrays.
binalgparser.c	Parser for binary alignments (e.g. water, needle, prophet). Module prefix bap_.
<b>blastparser.c</b>	Purpose: dissect the standard output of the BLAST program. Module prefix bp_.
chemutil.c	Chemistry tools: compound numbers, mol/sd files, InChIs. Module prefix chem_.
clientserverObject.c	TCP/IP client server functions. Module prefix: cso_.
<b>format.c</b>	dynamic String handling, C-string handling, Arrays of char*, line reading
graphalgo.c	Module containing algorithms for graph handling. Module prefix gral_.
hierclus.c	Hierarchical clustering of a square matrix with elements in the range [0.0..DBL_MAX]. Module prefix hc_.
<b>html.c</b>	Parsing HTML CGI POST data, various other CGI routines and generating HTML pages. Module prefixes cgi_, html_.
identifier.c	Knows how to verify the identity of a user. Module prefix ident_.
linestream.c	Module for reading arbitrarily long lines from files, pipes or buffers. Module prefix ls_.
pearsonfastaparser.c	Parser for output of Pearson fasta programs. Module prefix pfp_.
phraplightparser.c	Purpose: dissect the output of the PHRAP program in .ace files. Module prefix phrlp_.
plabla.c	PLatform ABstraction LAYer for Bioinformatics Objects and Services. Module prefix plabla_.
recipes.c	Routines from Numerical Recipes, Module prefix rcp_.
sequenceAlignment.c	to print an alignment between 2 sequences in a formatted fashion. Module prefix sa_.
sim4parser.c	dissect the output of the SIM4 program. Module prefix sim4p_.

# From the Bioinfo-C Examples: parsing BLAST output

```
#include "blastparser.h"  
#include "linestream.h"
```

```
static int subjectStart (char *subject, int subLen, char *desc) {  
    printf ("subjectStart: subject=%s len=%d\n", subject, subLen);  
    printf ("                desc=%s\n", desc);  
    numHSP = 0;  
    return 1;  
}
```

```
static int subjectEnd (void) {  
    printf ("subjectEnd: numHSP=%d\n", numHSP);  
    return 1;  
}
```

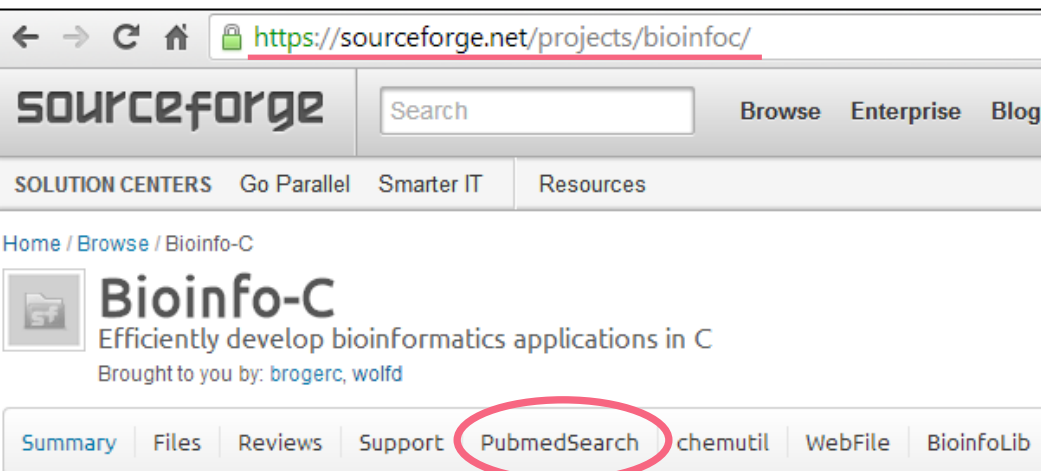
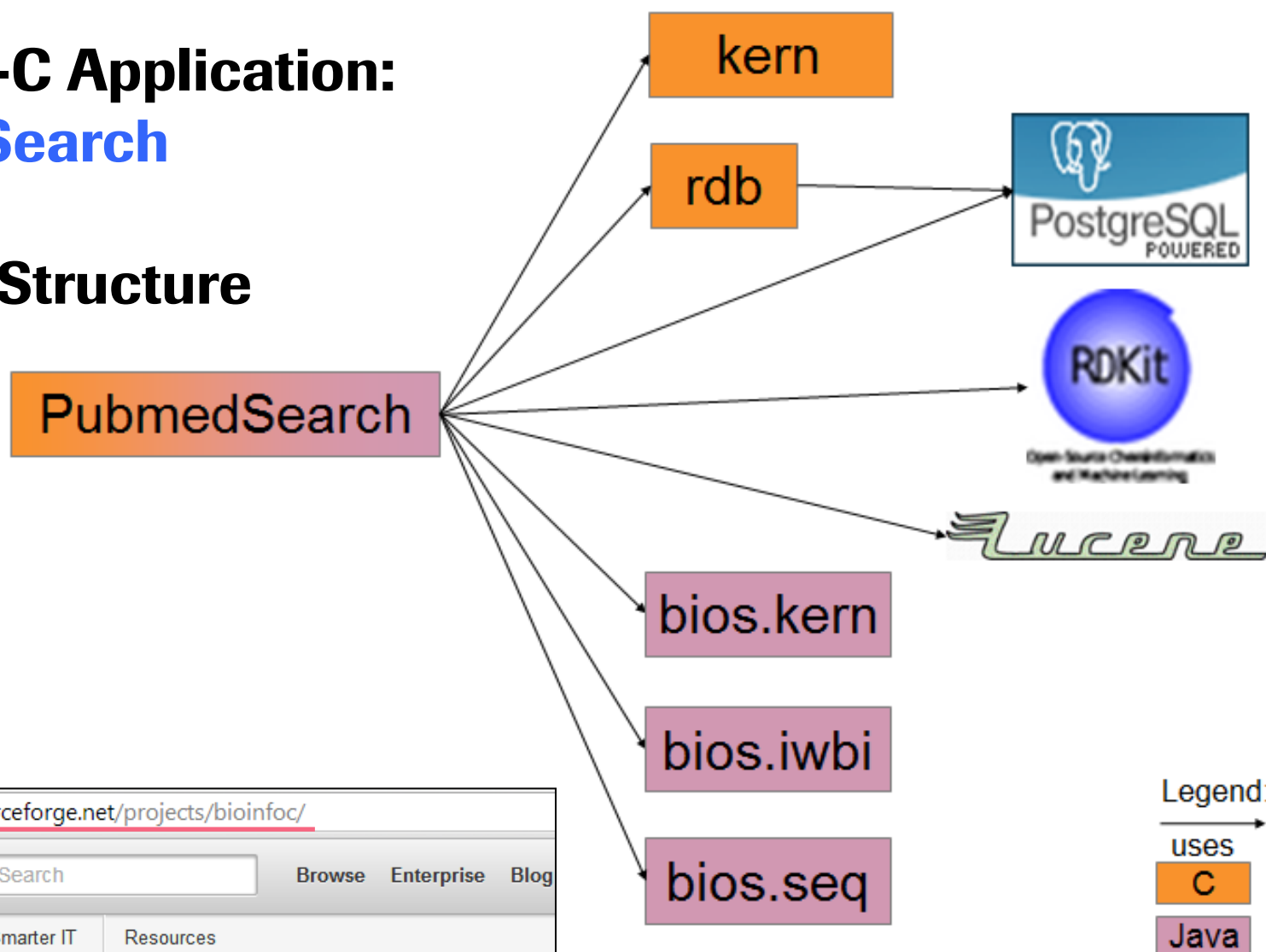
```
static int HSPStart (float score, double expect, int n) {  
    printf ("HSPStart: score=%f expect=%e n=%d\n",  
            score, expect, n);  
    numHSP++;  
    return 1;  
}
```

```
int main (int argc, char *argv[]) {  
    LineStream ls;  
  
    bp_init ();  
    bp_register_begin (&begin);  
    bp_register_end (&end);  
    bp_register_progName (&progName);  
    bp_register_query (&query);  
    bp_register_database (&database);  
    bp_register_summary (&summary);  
    bp_register_subjectStart (&subjectStart);  
    bp_register_subjectEnd (&subjectEnd);  
    bp_register_HSPStart (&HSPStart);  
    bp_register_idFrame (&frame);  
    bp_register_HSPEnd (&HSPEnd);  
    bp_register_HSPSeq (&HSPSeq);  
    ls = ls_createFromFile (argv[1]);  
    bp_run (ls);  
    ls_destroy (ls);  
    return 0;  
}
```



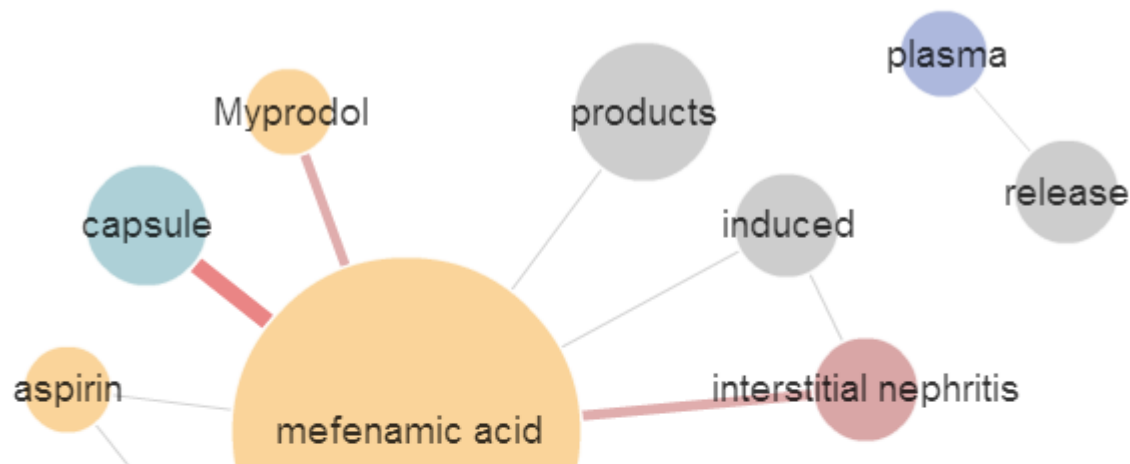
# A Bioinfo-C Application: Pubmed Search

## Software Structure



# Pubmed Search Example Output

Clemens Broger  
Yuan Wang

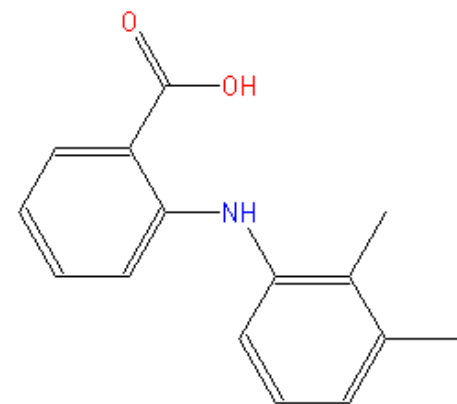


Found 17 document(s) in 6 milliseconds.

Create Object Summary

[Show parameters](#)

Document 1	<a href="#">Pubmed:24325782</a>	Date: 06/01/2014
Title	Aminocyanation by the Addition of N-CN Bonds to Arynes: Chemoselective <b>Synthesis</b> of 1,2-Bifunctional Aminobenzonitriles.	
Abstract	An efficient aminocyanation by the direct addition of aryl cyanamides to arynes is described, enabling incorporation of highly useful amino and cyano groups synchronously via <b>cleavage</b> of inert N-CN bonds, affording <b>synthetically</b> useful 1,2-bifunctional aminobenzonitriles. The postsynthetic <b>functionalization</b> of the aminocyanation <b>products</b> allows diverse formation of <b>synthetically</b> important derivatives such as drug molecule <b>Ponstan</b> and <b>fused</b> heterocycles.	
Author	Rao B, Zeng X	
Affiliation	Center for Organic Chemistry, Frontier Institute of Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, 710054, P. R. China.	
Journal	Org. Lett. 16:314-7(2014)	



Legend:

**term(s) matching query**

**Drug**

**Gene**

**GO**

**GeneGroup**

**Disease**

**Company/Institute**

Document 2	<a href="#">Pubmed:23426587</a>	Date: 21/02/2013
Title	[Infection after dental intervention. Iatrogenic or general medical <b>cause</b> ? Case report].	
	Whenever a dentist is dealing with <b>abscess</b> formation in the oral and maxillofacial region, it	

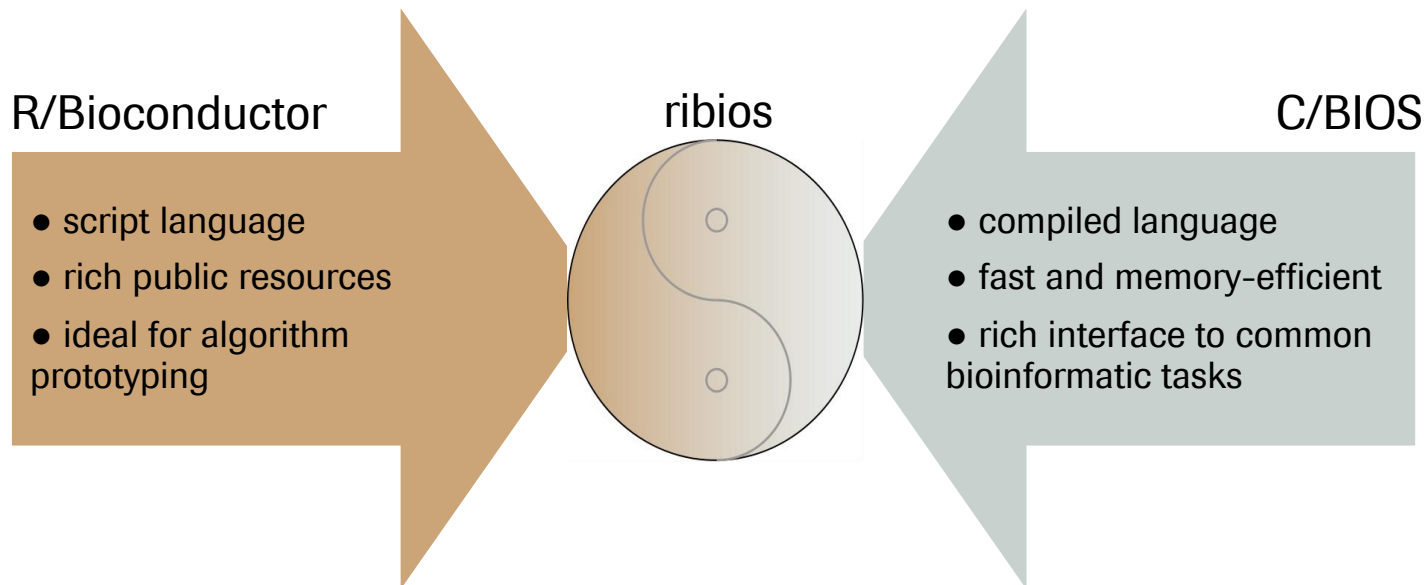
# ribios: Interfacing R and the BioinfoLib system

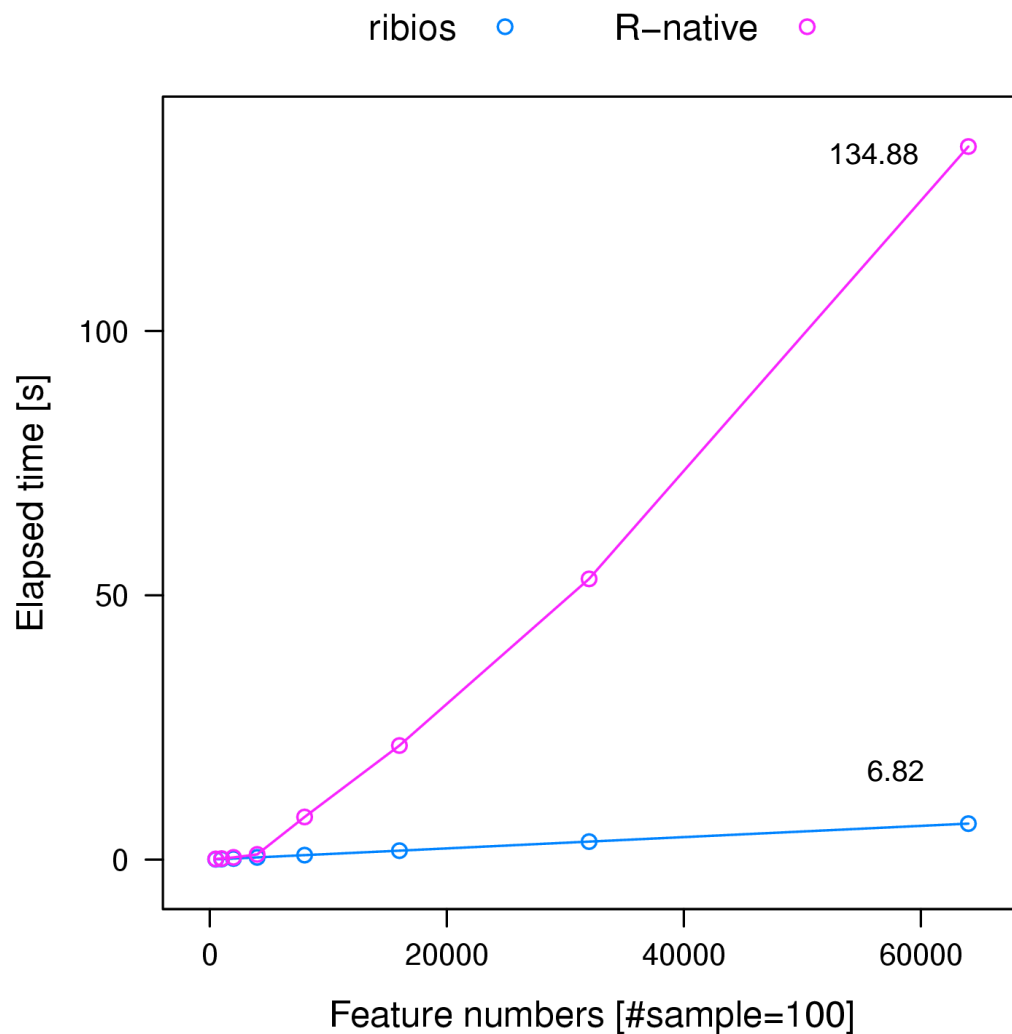


For bioinformatics tasks and interactive tools in computational biology

Jitao David Zhang

- ribios is a collection of R packages for **bioinformatics tasks** and **interactive tools** in **computational biology**
- It is an interface between R and the BioinfoLib/Bioinfo-C system (written in C),





**ribios example:  
reading a gene  
expression file**

**IO performance**  
*15~20 fold increase*

# Detlef Wolf

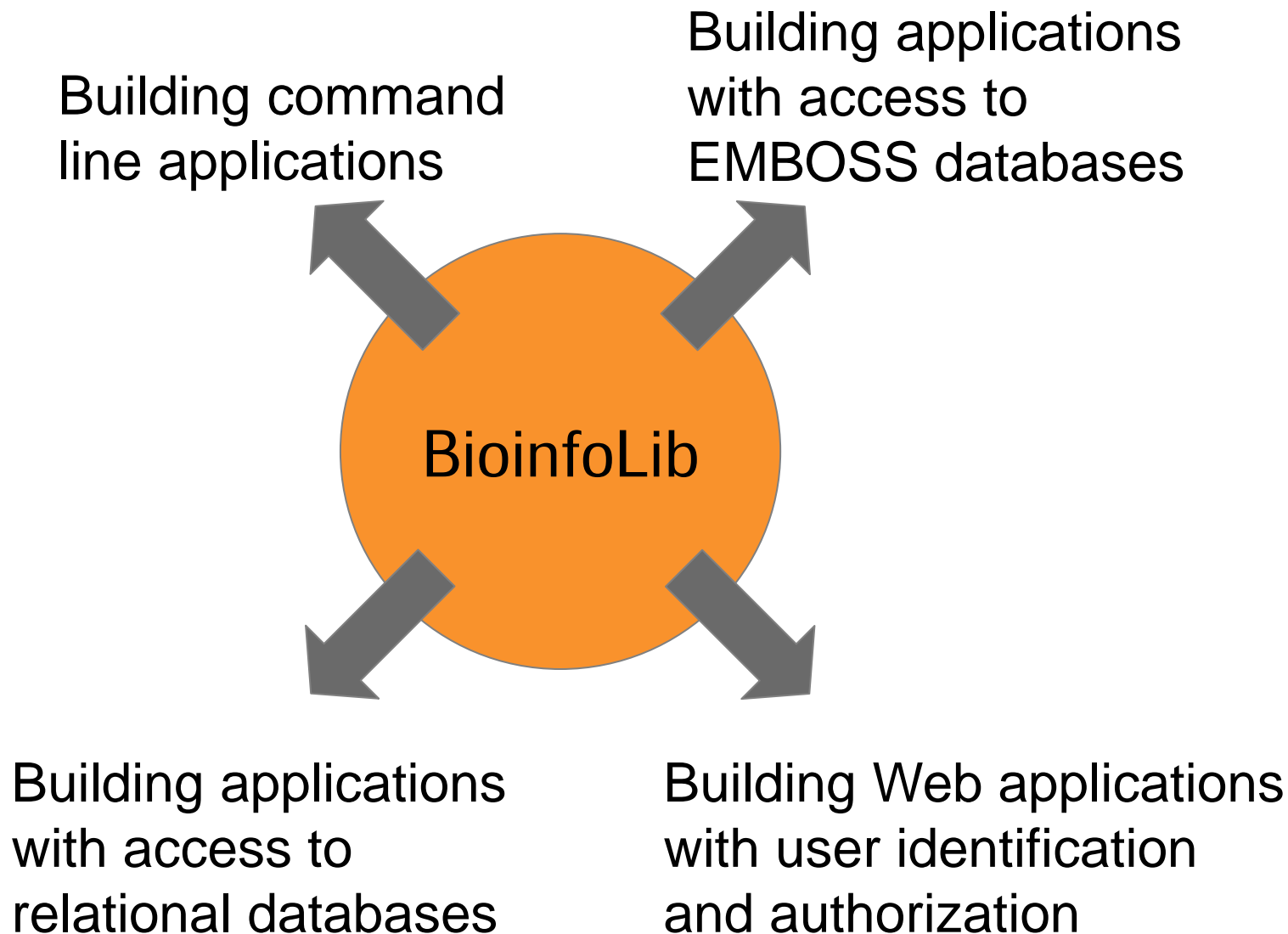
## Interests for this meeting:

- Roche Pharma Research is increasingly using open source software: Learn best practices for contributing and receiving.
- Learn how to turn BioinfoLib.kern (and possibly more) into a Debian package(s). BioinfoLib.kern enables object based C programming.

*Doing now what patients need next*

# **Backup Slides**

# Bioinfo-C Use Cases





# BioinfoLib kern module list

Module	Purpose
abigif.c	Creating gif file from an abi trace file. Module prefix abigif_.
affyfileHandler.c	Knows how to parse several kinds of Affymetrix files. Module prefixes cfo_ , ifo_ , ifo_ , dfo_ , efo_ .
algotil.c	Makes local or global sequence alignments. Uses code from EMBOSS. Module prefix algotil_.
alphatrans.c	Encode/decode byte stream for transmission over channel. Supporting only a restricted alphabet. Module prefix altr_.
arg.c	Module to parse command line arguments. Module prefix arg_.
array.c	Module for handling dynamic arrays.
avlTree.c	Module dealing with AVL balanced binary tree functions. Module prefix avl_.
binalgparser.c	Parser for binary alignments (e.g. water, needle, prophet). Module prefix bap_.
bitmap.c	Module dealing with bitmap objects. Module prefix bm_.
biurl.c	Knows how to map bioinformatics object identifiers to URLs.
blastdb.c	Reading sequences from blast databases. Module prefix bdb_.
blastparser.c	Purpose: dissect the standard output of the BLAST program. Module prefix bp_.
blastrun.c	Knows how to start the BLAST program and make its output available. Module prefix br_.
chemutil.c	Chemistry tools: compound numbers, mol/sd files, InChIs. Module prefix chem_.
chtoken.c	Dissect a C header file into tokens. Module prefix cht_.
clientserverObject.c	TCP/IP client server functions. Module prefix: cso_.
combi.c	Combinatorics functions. Module prefix cmb_.
dbsynonym.c	Determine sequence name synonyms for EMBOSS. Module prefix dsy_.
eval.c	Module that evaluates mathematical expressions. Module prefix eval_.
fastgenewise.c	Module to speed up genewise by shortening introns. Module prefix fg_.
forest.c	Module for representing hierarchical trees in main memory and printing them as an HTML selector. Module prefix forest_.
format.c	dynamic String handling, C-string handling, Arrays of char*, line reading
fuzzparser.c	Parses output of EMBOSS fuzz (nuc,pro,trans) programs. Module prefix fp_.
geometry.c	Various geometrical calculations Module prefix geom_.
gif.c	Creates GIF graphics. With the help of Die grosse Welt der Grafikformate Grafikprogrammierung unter Windows und WindowNT Thomas W. Lipp Synergy book ISBN 3-9803718-0-8. Module prefix gif_.
graphalgo.c	Module containing algorithms for graph handling. Module prefix gral_.
graphics.c	Drawing routines. Many of the routines come originally from the contribs of Daylight (www.daylight.com). Module prefix gr_.
hash.c	Hash table routines. Module prefix hash_.
hierclus.c	Hierarchical clustering of a square matrix with elements in the range [0.0,DBL_MAX]. Module prefix hc_.
hlrclock.c	Minimal and accurate time measurement. Module prefix hlr_.
hlrmisc.c	Miscellaneous nice routines not worth making separate modules from. Module prefix hlr_.
hmmparser.c	Purpose: dissect the output of the HMMSCAN or HMMSEARCH programs. Module prefix hmmp_.
html.c	Parsing HTML CGI POST data, various other CGI routines and generating HTML pages. Module prefixes cgi_ , html_.
htmlfile.c	Handle files coming from an HTML form. Module prefix htmlform_.
htmlform.c	Modification of HTML forms. Module prefix htmlform_.
http.c	HTTP GET and POST stuff: core communication routines of a WWW browser. Module prefix cgi_.
identifier.c	Knows how to verify the identity of a user. Module prefix ident_.
linestream.c	Module for reading arbitrarily long lines from files, pipes or buffers. Module prefix ls_.
Ink.c	Knows how to find bioinformatics object identifiers in natural language text and hyperlink them using module biurl. Module prefix Ink_.
log.c	Module for handling warnings, errors, etc.
mail.c	Send electronic mail. Module prefix mail_.
matvec.c	Purpose: basic routines for matrix and vector operations. Module prefix mv_.
msfparser.c	MSF (GCG multiple sequence file) parser, seems to work for ClustaW output as well. Module prefix msfp_.
notifier.c	Client for the event notification service: "Automatic Event Notifier". Module prefix notifier_.
pagedesign.c	Look and feel of bioinfoc.ch website. Module prefix pd_.
patternmatch.c	Simple pattern match routines. Module prefix pm_.
pearsonfastaparser.c	Parser for output of Pearson fasta programs. Module prefix pfp_.
phraplightparser.c	Purpose: dissect the output of the PHRAP program in .ace files. Module prefix phrlp_.
plabla.c	Platform ABstraction LAyer for Bioinformatics Objects and Services. Module prefix plabla_.
primer3parser.c	Parses output of the primer3 program. Module prefix pr3p_.
properties.c	the property file format understood by this C module is identical to the one used by the standard Java class java.util.Properties. Module prefix pty_.
pwddecode.c	Module 'encode/decode via C-string constant'. Module prefix endec_.
rdbu.c	Relational database utilities, handling database login info. Module prefix rdbu_.
rds.c	Provides access to user information from usrman. Module prefix rds_.
recipes.c	Routines from Numerical Recipes, The art of scientific computing, Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T., Cambridge University Press 1986. Algorithms have been adapted to real C (arrays start at 0). Module prefix rcp_.
regularexpression.c	Module to handle regular expressions. Examples for regular expressions: "^seq1 = ([^;]+).([0-9]+)bp\$". Module prefix regex_.
rotutil.c	Purpose: file handling utilities (and related stuff). Module prefixes: hlr_ , dio_.
seqautil.c	Sequence analysis utilities, promoter utilities. Module prefixes: seqa_ , prom_.
seqspeclst.c	Module for handling segmented list files as used in EMBOSS, not EMOSS dependent. Note: in more complex cases, use the sequenceObject/sequenceContainer. Module prefix seqspec_.
sequenceAlignment.c	to print an alignment between 2 sequences in a formatted fashion. Module prefix sa_.
sequtil.c	Module containing sequence utilities. Module prefix su_.
sim4parser.c	dissect the output of the SIM4 program. Module prefix sim4p_.
sim4run.c	Module that can run sim4. Module prefix s4r_ , M o
statistics.c	containing statistical algorithms from various sources. Module prefix stat_.
stringlist.c	associative arrayssociative arrays (i.e. arrays indexed by string)
useridprocessor.c	usrman user identifiers to user names entered in the form "firstname lastname". Module prefix uip_.
wiseparser.c	to parse genewise output. Module prefix wp_.
wwwsession	WWW sessions for CGI programs. Module prefix wwwsess_.
xmlbuilder.c	build string in XML format. Module prefix xmlb_.
xmlparser.c	Quick and Dirty event-based xml parser. Module prefix xmlp_.