Interactive primer- and probe-design with CaSSiS

Designing primers and probes for molecular diagnostic methods depends on the identification of oligonucleotide signatures, short binding sites on genome or marker gene sequences. CaSSiS is able to determine such signatures even under relaxed search conditions.

Kai Christian Bader and Harald Meier

Chair of Computer Architecture, Department of Informatics,
Technische Universität München, Boltzmannstraße 3, 85748 Garching

Bipartite Graph Representation Tree

CaSSiS relies on the Bipartite Graph Representation Tree (BGRT) [Bader et al., 2011], a newly developed data structure for fast sequence-to-signature mapping. Results are sorted by their degree of specificity, and all signatures guarantee a defined weighted mismatch value [Yilmaz et al., 2008] as a measurement for the Hamming distance to non-target sequences.

Comprehensive signature computation

CaSSiS' command line version was designed for the computation of comprehensive signature collections from large hierarchically clustered sequence datasets, i.e. signature candidates for every group and sequence. It aims at maintainers of sequence databases wishing to provide signature collections along with their public datasets.

Tested gene datasets

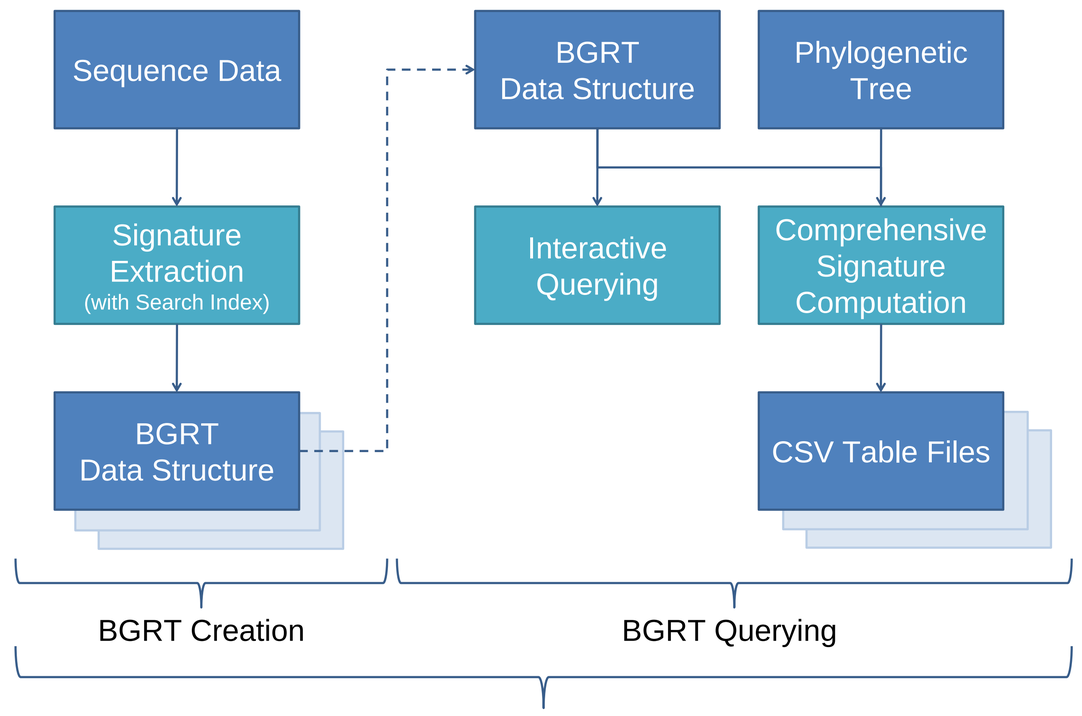
CaSSiS was successfully tested with SILVA databases [Pruesse et al., 2007], the largest collection of annotated aligned SSU-rRNA sequences of almost full length (>900 nt). The release *slv_106_red* (325,626 sequences, resulting in over 470 million relations) was processed on a Core i7 system within 20 hours with a peak memory consumption of 11 GBytes. Specific 18-mer signatures with full coverage were found for 15.9% of the corresponding phylogenetic groups and 71.8% of the sequences. Allowing up to 10 non-target matches increased their number to 39.0% and 92.1%.

Signature extraction & storage

For end users primarily interested in few defined target groups (but still with large background data), single queries instead of a comprehensive signature set computation are sufficient. CaSSiS was therefore extended: Computationally intensive mappings on high-performance systems can now be stored. (The *slv_106_red* BGRT needs only 856 MBytes of space.) End users can then use these mappings for querying user-defined groups.

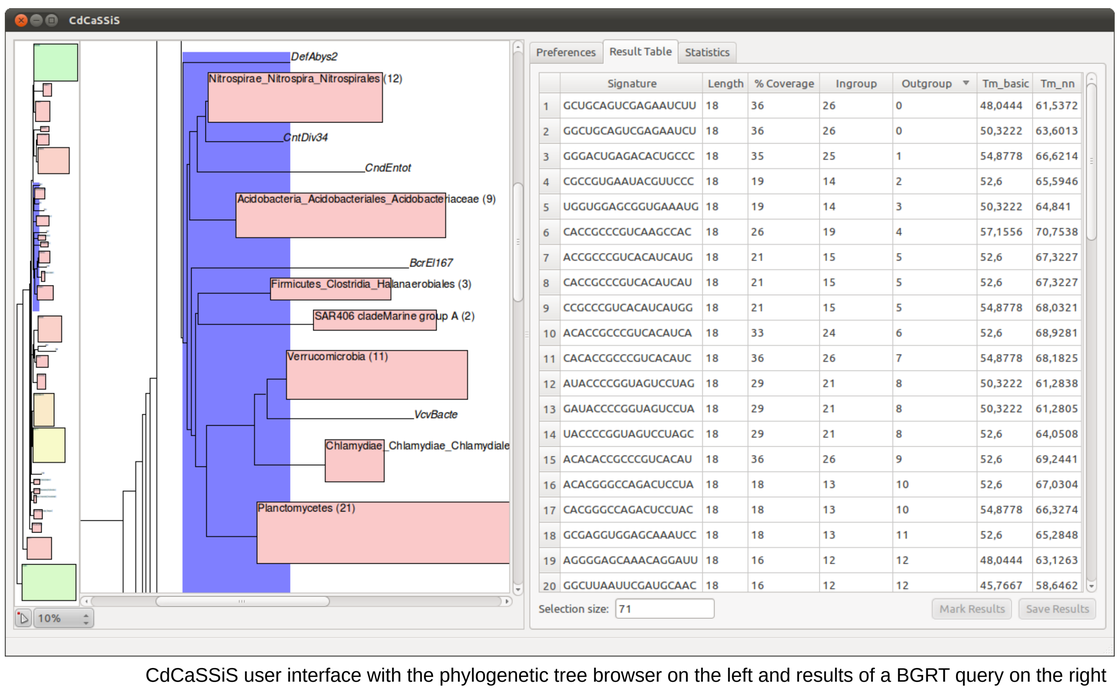
Interactive querying

A new, intuitive graphical interface allows group selection within a loaded phylogenetic tree or its definition as a list of identifiers. By allowing a range of non-target hits, users may influence the specificity. Single requests are usually processed within a second. The result is a list of signatures with maximum coverage (sensitivity) for each entry within the range of allowed non-target matches and their thermodynamic characteristics.



	A	B	C	D	E	F	G
1	Index	Species/Group ID	Size	Ingroup	Signatures...		
2	2	Bacteria	278862	248618	CCUACGGGAGGCAGCAGU		
3	582330	Eukarya	34451	31030	GUGGUGCAUGGCCGUUCU		
4	582359	Opisthokonta	20523	4862	GCGCGCUACACUGAAGGA		
5	582362	Metazoa	12246	4862	GCGCGCUACACUGAAGGA		
6	571775	Crenarchaeota	5192	4361	UGGUGUCAGCCGCCGCGG		
7	557725	Archaea	12303	4246	CGGUGCCAGCCGCCGCGG		
8	557731	Euryarchaeota	7013	4133	ACCGGUGCCAGCCGCCGC		
9	624663	Archaeplastida	4324	3958	AGAACGAAAGUUGGGGGC		
10	473823	Actinobacteria_Actinobacteria	26798	3717	GCCGGUGGCCCAACCCCU		
11	16	Proteobacteria	98192	3668	AGAGUAUGGUAGAGGGUG		
12	473832	Actinobacteridae	22838	3542	GCGACAUUCCACGUCGUC		
13	411362	Firmicutes_Bacilli	22697	3166	GGGUCAUUGGAAACUGGA		
14	582378	Arthropoda	6088	3104	CGGGACUCAUCCGAGGCC		
15	607212	Fungi	8092	2614	GCUCAAGCCGAUGGAAGU		
16	34	Gammaproteobacteria_1	34345	2595	GGGGGUAGAAUUUCAGGU		
17	473835	Actinomycetales_1	11758	2544	GAGUUCGGUAGGGGAGAU		
18	68729	Betaproteobacteria	18283	2454	UGGCAGAGGGGGGUAGAA		
19	473849	Corynebacterineae	5449	2051	GGCGAUACGGGCAUAACU		
20	122908	Alphaproteobacteria	26902	1562	AGAAUUUCUAGUGUAGAG		
21	411388	Lactobacillales	8390	1494	AUCUUACCAGAAAGGGAC		
22	406458	Anaerococcus_1	1374	1201	UCAAAAAGCCUGUCCCAG	CUCAAAAAGCCUGUCCCA	
23	68743	Burkholderiales	13135	1018	GGCGAAAGCUUUGCUAAU	AGGCGAAAGCUUUGCUAA	AACGAGGCGAAAGCUUUG
24	607217	Ascomycota	5115	922	CCGUUCGGCACCUUACGA	CCCGUUCGGCACCUUACG	
25	607220	Ascomycota	5052	922	CCGUUCGGCACCUUACGA	CCCGUUCGGCACCUUACG	

Comprehensive signature computation result (CSV dataset, 0 mismatches, 1 non-target hit)



CdCaSSiS user interface with the phylogenetic tree browser on the left and results of a BGRT query on the right

Literature:

- Kai Christian Bader, Christian Grothoff, and Harald Meier. Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. Bioinformatics, 27: 1546-1554, June 2011. http://dx.doi.org/10.1093/bioinformatics/btr161
- Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M. Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank O. Glöckner. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res., 35 (21):7188-7196, December 2007. http://dx.doi.org/10.1093/nar/gkm864
- L. Safak Yilmaz, Lindsey I. Bergsven, and Daniel R. Noguera. Systematic evaluation of single mismatch stability predictors for fluorescence in situ hybridization. Environmental Microbiology, 10(10):2872-2885, 2008. ISSN 1462-2920. http://dx.doi.org/10.1111/j.1462-2920.2008.01719.x

Availability: CaSSiS is available for download at http://cassis.in.tum.de/

Contact: Kai Christian Bader (baderk@in.tum.de)

Funding: Bayerische Forschungsstiftung