GENtle, a free multi-purpose molecular biology tool

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Magnus Manske
aus
Köln

2006

Berichterstatter:	Prof. Dr. Helmut W. Klein
	Prof. Dr. Sabine Waffenschmidt

Tag der mündlichen Prüfung : 29. November 2006

MEINEN ELTERN

Die vorliegende Arbeit wurde in der Zeit von Januar 2003 bis September 2006 am Institut für Biochemie an der Universität zu Köln unter der Leitung von Herrn Prof. Dr. Helmut W. Klein angefertigt.

Table of Contents

Table of figures

1 Zusammenfassung

Als Resultat moderner molekularbiologischer Technologien, nicht zuletzt der DNA-Sequenzierung, wächst das Volumen biologischer Daten seit Jahren exponentiell. Neben großen, hochspezialisierten Datenbanken, die über das Internet zur Verfügung stehen, ergibt sich für alle Arbeitsgruppen die Notwendigkeit, Sequenzen zu analysieren, modifizieren und organisieren. Trotzdem leidet die dafür notwendige Software häufig an Problemen: Existierende, freie Software deckt oft nur einen kleinen Teil der notwendigen Funktionen ab, ist schwer zu installieren und zu bedienen; kommerzielle Software ist nicht selten auf eine Funktionsgruppe spezialisiert, und zwingt den Benutzer in proprietäre Formate.

Im Rahmen meiner Doktorarbeit habe ich GENtle entwickelt, ein freies Programm mit einem breiten Spektrum molekularbiologischer Funktionen für den täglichen Bedarf im Labor, die sich übergangslos in ein Paket integrieren. Dieses in C++ geschriebene Programm läuft auf mehreren Betriebssystemen, wurde auf Geschwindigkeit optimiert und stellt eine Datenbank-basierte Sequenzverwaltung zur Verfügung.

Die Funktionen von GENtle umfassen die Verwaltung, Bearbeitung und Analyse von DNA- und Aminosäuresequenzen, virtuelle Klonierung, Gele, PCR, Primer-Erstellung und -Optimierung, Erstellung und Layout von Alignments, Chromatogramm- und Gelbilddarstellung, sowie zahlreiche zugehörige Funktionen.

Ziel meiner Entwicklung war es, ein leistungsstarkes, breit angelegtes und dennoch einfach zu bedienendes System zu entwickeln. GENtle hat sich inzwischen nicht nur in unserer Arbeitsgruppe, sondern in Labors weltweit bewährt.

2 Abstract

A result of modern techniques in molecular biology, especially DNA sequencing, is the exponentially growing amount of available data. Besides giant, specialized databases, which are accessible over the Internet, all work groups in the field of molecular biology today need to handle, modify, analyze and store sequence information. This trend notwithstanding, general purpose software for these tasks often suffers from severe drawbacks. Free software exists, but is often hard to set up and operate for users on today's point-and-click interfaces, and usually leads to the application of a patch-work of multiple, only partially compatible tools and web services. Commercial software often covers only parts of the required functions, and tends to lock the user into proprietary formats.

In my thesis, I have developed GENtle, a free, multi-purpose bioinformatics software, seamlessly integrating diverse applications for every-day lab use in a single package. It was designed for easy and intuitive use, while providing many powerful functions. This C++ application runs on multiple platforms, is optimized for performance, and includes database interfaces for easy sequence management. It features DNA and protein sequence management and analysis, virtual cloning, gels, and PCR, primer design, alignment generation and layout, chromatogram and image display, as well as many related functions. GENtle strives to satisfy the need for an easy and comfortable, yet powerful multi-purpose tool.

One design goal of GENtle was "instant responsiveness". Likewise, consistent display and handling are of great importance. GENtle has been outfitted with modules for DNA and protein sequence management, editing, and analysis, primer design, virtual PCR, alignments, virtual gels, a plethora of import and export formats, integrated database management, internet search functionality, an auto-update mechanism, and a number of integrated tools. GENtle is free software licensed under the GPL and available for Windows, Mac OSX, and Linux in several languages. As such, it is already in use in research groups worldwide.

3 Introduction

3.1 Bioinformatics in the lab

Bioinformatics, also called computational biology, uses methods from informatics, statistics, computer science, and applied mathematics to work on problems of molecular biology, especially biochemistry and genetics. Once a new domain restricted to a few experts, applications of bioinformatics today have penetrated into the daily routine of every work group in the field of molecular biology. Searching in online sequence databases have become a vital tool as much as three-dimensional protein structure visualization, computer-aided primer design, and sequence analysis, annotation, and management.

3.1.1 Information overload

In former times, the major problem of information was its lack. Only a few decades ago, publications were rare commodities, and even a well-stocked library only carried a fraction of the available material. Since the advent of applied information theory in the shape of the (personal) computer, this issue has essentially been reversed. While finding the information you want is still the task, it is now difficult not because of the absence of information, but because of the abundance thereof. This so-called "information overload" not only affects publications, with PubMed entries currently increasing at ca. 3000 articles per day, but also raw data repositories. The number of known DNA sequences and genomes increases at an exponential rate (Figure 1). An army of machines, ranging from DNA sequencers, FPLCs, HPLCs, and photometers to mass spectrometers and microarray readers, produce digital data in amounts never before imagined. This data

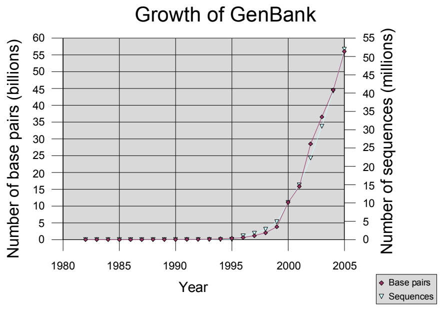


Figure 1: Growth of GenBank.

The number of base pairs and DNA sequences in GenBank. Source : http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html has to be stored, analyzed, compared, referenced, summarized, and made available to the public.

With both increasing sequence data available and improved tools for creating variants thereof (restriction endonucleases, PCR, etc.), research groups tend to create more sequence data, and therefore increase the need to organize this information.

3.2 GENtle, the Bad, and the Ugly

While the needs of different work groups for bioinformatics tools obviously vary, many laboratories share some basic necessities, most notably the management, display, manipulation, and analysis of protein and nucleic acid sequences. The manipulation of DNA, including restriction endonuclease cleavage, ligase joining, and polymerase chain reaction, are universal operations in all work groups handling DNA.

Predictably, numerous software exists to address the problems arising from these operations. However, I found none of the existing programs to be satisfactory for our work group.

It has been a long-standing tradition in science to freely share research, thus facilitating the accumulation of knowledge, as each researcher can “stand on the shoulders of giants” (Bernhard von Chartres, ca. 1130; often attributed to Newton, 1675). Today, the shared information encompasses results, methods, and data.

But how can data, for example sequence information, be called “free” when the software required needs to be purchased? As with most hardware specifically designed for laboratory use, prices for commercial software are quite significant for all but the most well funded work groups, easily ranging into the hundreds of thousands of Euros for laboratory-wide installations. Not to be ignored either are the costs of updates/upgrades, which are often mandatory. Also, commercial software usually covers only a single area of functions; for example, a sequence management package often lacks a module for primer design. Last not least, commercial software tends to lock data into a proprietary format, aggravating a switch to another software at a later date (*lock-in*).

Free and open source software (FOSS) has a long tradition in scientific computing [1]. Consequently, several software packages exist. One of the most mature packages is EMBOSS [2], offering hundreds of functions for sequence analysis, search, and prediction. Its structure is based on the UNIX philosophy, which is the origin of bioinformatics software, and is thus comprised of many small programs, each of which performing a single function only, and intended for use on the command line. While graphic interfaces exist, setup and handling of the EMBOSS package can prove difficult for researchers not intimately familiar with either UNIX or bioinformatics.

There is a plethora of free bioinformatics tools available, both as local or web applications. However, most of these applications are severely limited in their functional spectrum, thus lacking a consistent handling. Also, sequences have to be imported and exported repeatedly, potentially leading to loss of meta-data such as annotations and comments; even more so in the case of web applications, which usually do not accept annotations at all. Last not least, most web applications do not support encrypted data transmission (e.g., https), thus posing a potential security threat, as sensitive sequences are transmitted over the internet for everyone to read.

In face of these drawbacks, I decided to address these issues by developing my own software package. GENtle grew on practical requirements in our work group as much as on existing examples and solutions to typical problems in the field.

3.3 Aim of the project

In this project, my goal is to provide a free software for the admittedly huge field of molecular biology, focusing on but not limited to sequence display, analysis, and manipulation, that is easy to set up and use, while providing practical and powerful functions.

3.3.1 Usability

An essential but often overlooked part of any computer program is the human-computer interaction (HCI) [3] via the user interface. Therefore, a major concern of mine was the rather abstract concept of “usability“. In 1991, the usability of a software product was defined by ISO 9126 (http://www.iso.org) as “a set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users“. This definition was extended in 1998 by ISO 9241-11 as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use“.

In this concrete project, usability has multiple facets. One is a low entrance barrier, that is, the need for the user to learn how to work with that unfamiliar software. A guiding principle is that of the “least surprise” [4] - the user should find data and functions where he expects them (among interface designers colloquially also known as KISS, “keep it simple, stupid”). GENtle uses a single design (often called “look and feel”) for all its modules, so identical elements are recognizable across different modules. Data that belongs together is displayed together; restriction endonuclease sites are shown directly below the DNA sequence in place; similarly, annotations and the amino acid sequence that would result from the DNA in the current reading frame are shown above the DNA sequence. In the PCR module, primers are displayed next to the appropriate template DNA sequence. Protease cuts are shown within the amino acid sequence. All sequences can be edited in place, and the surrounding information changes in tune with the sequence.

Program functions have been modeled after their biological counterpart as much as possible, mimicking the user's known working steps *in silico*. The user designs virtual primers based on a DNA sequence, then runs a virtual PCR, cuts DNA sequences with virtual restriction endonucleases, optionally generating a virtual agarose gel in the process, then uses virtual ligation to connect matching ends. The resulting DNA can be virtually translated into an amino acid sequence. These processes are modeled after the background one can expect from a potential user, thus giving that user a “feeling of familiarity” and further lowering the entrance barrier.

Also, consistency is an important aspect of usability. Annotations and enzyme sets are maintained together with a sequence when moving from one module to another, even when translating DNA to amino acids (and back).

Another usability factor is speed. If a user has to wait for a computer program, his or her work flow is interrupted, which can have a significant negative impact on the efficiency with which the user can operate. Even brief delays, if they only occur often enough, can add to the dissatisfaction of the user with the software. Consequently, I designed GENtle to run as smoothly as possible, even on machines that are somewhat out of date. This was also a reason which made me rule out programming languages depending on virtual machines like Java and C#, and choosing C/C++ instead.

Simple installation and an integrated update mechanism, context-sensitive online help as a wiki, extensive use of context menus, and an easy-to-use database interface are further key points of my usability efforts, letting the user concentrate on his work instead of how to work the software.

3.3.2 Function spectrum

There is a plethora of tools and functions available in bioinformatics, and any attempt to cover all of them in a single piece of software is doomed by default. I decided that GENtle should consist of a set of basic modules, and be extended according to need in the reality of the lab.

Consequently, the sequence display, editing and analysis modules comprise the core of GENtle. DNA and amino acid manipulation, alignments, and primer design and PCR, likely comprise the majority of every-day work in most laboratories concerned with molecular biology.

In addition to these basic modules, GENtle offers a variety of functions that have been requested by users of my work group, by email, or even by users of other software. These include modules to generate virtual gels, to query online sequence databases, and run BLAST searches, but also functions one would not expect, like an image viewer for obscure gel image formats, translating amino acid sequences back into DNA with a given species preference, or a function to read a sequence out loud. The software has been adapted and extended time and again to fit the needs of the user, so the user would not have to fit to the needs of the software.

3.4 Availability of GENtle

GENtle is available for Windows (Windows 95 or above), Mac OSX (10.2 or above; native on PowerPC, using ROSETTA for Intel processors) and Linux (x86 binary) at http://gentle.magnusmanske.de. The source code, including XCODE configuration and Linux AUTOCONF files is available via CONCURRENT VERSIONING SYSTEM (CVS) following the instructions at http://sourceforge.net/cvs/?group_id=89980. A DOXYGEN code documentation is available at http://gentle.magnusmanske.de/doxygen/annotated.html. An online manual is available at http://en.wikibooks.org/wiki/GENtle_manual.

4 Materials and Methods

4.1 General equipment

For the development of the different versions of GENtle, the respective platforms were used. Windows development took place on Windows 2000 and XP, the Mac version was developed under OSX 10.4 on a PowerBook G4 (PowerPC), and the Linux version was compiled under the UBUNTU OS 6.06 LTS *Drapper Drake*. All systems were 32-bit, though I was told about successful compilations on 64-bit Linux systems.

4.2 Software

Only free (as in „free beer“) software was used in the development of GENtle and its implementation on the supported platforms. Most of it is public domain or free (as in „freedom“) software, available under licenses such as the *GNU General Public License* (GPL), available at http://www.gnu.org/licenses/gpl.html. A list of open source licenses can be found at http://www.opensource.org/.

For Windows and Max OSX, integrated development environments (IDEs) were used. The Linux version is compiled directly from command line, using the AUTOCONF/AUTOMAKE tool set.

4.2.1 Dev-C++

DEV-C++ is an integrated C/C++ programming language development environment based on the MINGW compiler, which is a Windows port of the GCC (GNU COMPILER COLLECTION). The IDE itself is written in the Delphi programming language. Both are licensed under the GPL. I found DEV-C++ useful for the task of developing GENtle for Windows, as it provides a suitable editor, an integrated compiler, and the ability to use special extension packages („DevPacks“) available from various sources on the WWW, especially a precompiled wxWidgets package. DEV-C++ is available at http://www.bloodshed.net/devcpp.html. For development, DEV-C++ version 5 (beta) was used.

4.2.2 Xcode

XCODE is an IDE for Mac OSX, made available by Apple at no cost, and widely used among Mac programmers. Several versions of XCODE were used during the Mac development of GENtle. An overview of XCODE is available at http://www.apple.com/de/macosx/features/xcode/, though free registration as a Mac developer is required to download and use it.

4.2.3 wxWidgets

WXWIDGETS (formerly WXWINDOWS) is a cross-platform toolkit, allowing me to write code that will compile and run on several different operating systems, provided it is linked to the respective library on that OS. I am using WXWIDGETS on Windows (WXMSW), Mac (WXMAC), and Linux (WXGTK), though it is available for more platforms.

Despite its name, WXWIDGETS not only provides visual components such as windows, menus, and buttons, but also offers platform-independent access to file system and network, right down to its own array and string classes, which have proven useful for their Unicode support. WXWIDGETS is available under LGPL license at http://wxwidgets.org/.

4.2.4 sqlite

SQLITE is a database engine that realizes a complete database in a single file. In contrast to other database systems, no database server is needed; SQLITE is usually linked to an application, as I have done with GENtle, using both version 3 (SQLITE3) by default and version 2 (SQLITE2) for downward compatability. As the name suggests, SQLITE supports most of the SQL92 standard, allowing me to use the same SQL commands for both MYSQL and SQLITE. I chose sqlite as the default database engine, as it relieves the user of the necessity to set up a MYSQL server, especially for testing and single-workplace installations. The SQLITE code is in the public domain. Source, libraries, and binaries are available at http://sqlite.org/.

4.2.5 MySQL

MySQL is a widely used and powerful database engine. It is able to support databases of virtually unlimited size, equaling purely commercial engines such as ORACLE or MSSQL. I am using MYSQL as an alternative database engine in GENtle to support sharing of sequence information within work groups. Apart from the setup, usage of MYSQL and sqlite database does not differ from the user perspective. MYSQL is available under the GPL at http://mysql.com/.

4.2.6 TinyXML

TINYXML is a collection of C++ files that comprise a XML parser, turning XML text into a structured, tree-like internal representation. I am using TINYXML in GENtle to parse files in XML format, e.g., GenBankXML. It is also used internally for some sequence descriptions (e.g., alignments) in the database. TINYXML is free software under the zlib license and available at http://www.grinninglizard.com/tinyxml/.

4.2.7 NSIS

The NULLSOFT SCRIPTABLE INSTALL SYSTEM (NSIS) is a software to create Windows installers, that is, compressed, self-extracting archives that provide an installation dialog, uncompress and copy included files to a chosen location, and register the installed application with the Windows registry. The installation process can be defined though a simple scripting language, hence the name. I chose NSIS because it is easy to use for the end user, has a powerful scripting language, and supports LZMA compression, one of the best compression algorithms for executable files, yielding a compression factor of more than 4 for GENtle and associated files. NSIS is available under a free license at no cost under http://nsis.sourceforge.net/.

4.2.8 Doxygen

DOXYGEN is a code documentation system. It works by analyzing specially formatted comments within the program source, and generating an annotated, structured, and interlinked description of global functions and variables, classes, methods, and member variables. While several output formats are available, I only used the HTML output due to its flexibility. DOXYGEN is licensed under the GPL, and available at http://www.stack.nl/~dimitri/doxygen/. A recent DOXYGEN HTML output for the GENtle source code is available at http://gentle.magnusmanske.de/doxygen/annotated.html.

4.3 Algorithms

For GENtle, I have used several algorithms well-known in bioinformatics, using existing C/C++ code where suitable, and writing the code myself in other cases.

4.3.1 Alignments

GENtle comes with three built-in alignment algorithms.

4.3.1.1 Needleman-Wunsch

The Needleman-Wunsch algorithm [5] is a classic algorithm which performs a global alignment between two sequences, for both nucleotide and amino acid sequences. It defines an alignment score based on exact matches, mismatches, and gaps; the first one increases, the latter two decrease the score. The algorithm is guaranteed to find the alignment with the highest score.

The algorithm is based on a matrix, with a column for each letter of sequence A, and a row for each letter of sequence B. The resulting matrix F thus has the dimensions $n_A \times n_B$, and its elements are referred to as F_{ij}. The algorithm then calculates the alignment with the highest score. It recursively calculates $F_{ij} = max(F_{(i-1)\times(j-1)} + S(A_i, B_j) \times F_{i\times(j-1)} + d \times F_{(i-1)\times j} + d)$ with $S(a,b)$ being the (mis)match weight of the letters a and b and d being the gap penalty.

Once this matrix is generated, the alignment with the highest score can be backtracked starting from the lower right corner. For each element, there are three choices (routes to take): up, left, or diagonal up-left. The latter equals a match, the other two represent a gap in the respective sequence. The backtracking algorithm always follows the choice with the highest score. Depending on where the algorithm stops (the upper-left corner, somewhere in the first row, or somewhere in the first column), leading gaps have to be introduced into the respective sequence.

A practical problem with this algorithm is the size of the matrix, which can become quite large if long sequences are aligned. For this reason, the algorithm is often implemented using Hirschberg's algorithm for the longest common subsequence problem [6], buying less memory usage with reduced speed. As one of my design goals was high speed, I implemented this algorithm myself using a slightly altered system that allows for the matrix elements to be stored as single bits, allowing for a matrix with a reduced memory footprint while maintaining the speed of the original algorithm. The cost is the loss of fine-tuning in the mismatch weights S , which I consider negligible for the application of this type of algorithm within GENtle, namely quick comparisons and alignments to sequencer data.

My implementation also allows for a simple multi-sequence alignment, where all sequences are compared to the first one in a list of sequences. Gaps that are introduced in the first sequence are also introduced in all other sequences already aligned. This method is quick but prone to the erroneous insertion to gaps. Again, I consider this to be acceptable for the task at hand. High-quality alignments will be made using the advanced ClustalW algorithm.

4.3.1.2 Smith-Waterman

The Smith-Waterman algorithm [7] is a variation of the above Needleman-Wunsch algorithm, suited for local alignments. Based on $F_{11}=0, F_{1j}=0, F_{i1}=0$, it improves on its predecessor in cases where two sequences are very different except for a few conserved regions, or where one sequence should align to a small part of another, much longer sequence. This is achieved by setting negative values of the substitution matrix S to zero, widely ignoring mismatches. My implementation reuses my Needleman-Wunsch code with a few parameterized differences. I recommend this algorithm to compare sequencer data to the original sequence.

4.3.1.3 Clustal

The Clustal algorithm [8] is a multiple sequence alignment method. It works by creating a phylogenetic tree to group sequences by similarities, then using a substitution matrix to calculate the likelihood of exchanges (allowed mismatches), and finally fine-tuning gap opening and extension.

Two gap types exist, initial and position-specific gaps. The initial gap type depends on the weight matrix, as well as on the similarity and the length of sequences. Thus, the gap-opening penalty GOP is calculated as $(GOP_{wm} + \log(min(n,m))) * (\text{Ø}_{rm}) * (P_{isf})$ with GOP_{wm} being the gap opening penalty according to the weight matrix, m and n the respective length of the sequences, Ø_{rm} the average residue mismatch score, and P_{isf} the percent identity scaling factor. Likewise, the gap extension penalty GEP is calculated as $GEP_{wm} \times (1.0 + \left|(\log(\frac{n}{m}))\right|)$ with GEP_{wm} being the gap extension penalty according to the weight matrix.

Initial gap penalties can be modified as position-specific gap penalties. GOP and GEP penalties are modified according to position. For existing gaps, the GOP is reduced to $GOP_{initial} \times 0.3 \times \frac{\text{no. of sequences without a gap}}{\text{no. of sequences}}$, and near existing gaps, the GOP is increased by $GOP_{initial} \times (2 + \frac{8 - \text{distance from gap} \times 2}{8})$. Additionally, hydrophilic stretches of five residues reduce the GOP by $^1/_3$, and GOPs for non-gap stretches are multiplied with a weighting factor specific for the residue. This is intended primarily for amino acid sequences, but works just as well for nucleotide sequences using an appropriate weight matrix.

For GENtle, I am using the official ClustalW sources, which are written in C, directly as part of the program. The authors state ClustalW to be „free for academic users“. A `#define` switch can toggle between compiling the ClustalW source and using a standalone Clustal executable instead; I implemented this to be compatible with Linux

Debian-based distributions, which do not allow software to contain code that already exists as a tool in that distribution. Both ways yield identical results in GENtle.

4.3.2 Other algorithms

4.3.2.1 Ncoils

The COILS algorithm [9] attempts to find coiled-coil structures based on the amino acid sequence. I am using the source of the Ncoils program by Robert Russell, who released it under the GPL. http://www.russell.embl.de/cgi-bin/coils-svr.pl has both source and online tool.

4.3.2.2 siRNA

Small interfering RNA [10] refers to 20-25 nucleotide-long RNA molecules which can interfere with the expression of a gene containing that sequence. It works by iterating through a coding DNA sequence, analyzing windows with a length of 21 nucleotides, calculating window scores for different properties. I have re-implemented the algorithm of the EMBOSS siRNA tool at http://emboss.sourceforge.net/apps/cvs/sirna.html.

4.3.2.3 Isotopic Pattern Calculator

The Isotopic Pattern Calculator (IPC) is a software by Dirk Nolting, calculating the isotopic distribution of a chemical compound. I am using the algorithm in its original source to generate a mass spectrometer preview for peptides, based on their amino acid sequence. The display is achieved through the graph/spectra module. IPC is available under GPL at http://isotopatcalc.sourceforge.net/.

4.3.2.4 UReadSeq

UReadSeq is a set of import/export filters for various sequence formats, written by D. G. Gilbert. I am using the 1993 C code as a fall-back import filter, in case my self-written filters fail to recognize the file format. UReadSeq reads the name and sequence information only, removing any annotation in the process. It is public domain, available at http://iubio.bio.indiana.edu/soft/molbio/readseq/.

4.3.2.5 Chou-Fasman secondary structure prediction

The Chou-Fasman algorithm [11] is an algorithm that tries to predict the secondary structure (α helices, β sheets, turns) of a protein based on its amino acid sequence. The algorithm is based on a table of statistical probabilities of an amino acid being part of a secondary structure.

The algorithm itself is imprecise compared to more modern ones, which involve large database, pattern recognition, and sometimes neural networks. It is, however, very fast, can run locally, and does not depend on any database other than the amino acid table. It thus meets the standard of instantaneous responsiveness I have set for GENtle. Also, no other algorithm is able to guarantee a correct prediction either.

I have implemented this algorithm myself based on the probability table and description at http://prowl.rockefeller.edu/aainfo/chou.htm.

4.3.2.6 Hydrophobicity

The algorithm to predict local hydropathicity/hydrophobicity is using data and methods by Kyte-Doolittle [12] and Hopp-Woods [13], respectively.

4.4 Online resources and services

4.4.1 BLAST

Based on a protein or nucleotide sequence, the Basic local alignment search tool (BLAST) can be used to find similar sequences in a database [14]. In GENtle, either can be searched through the NCBI web service interface, using blastp for protein and blastn for nucleotide sequences, respectively. A search returns a preview alignment of the original and the putative sequence. The latter can be retrieved as a new sequence. The NCBI BLAST service is available at http://www.ncbi.nlm.nih.gov/blast/index.shtml.

4.4.2 PubMed

PubMed is a database by the U.S. National Library of Medicine, well known to contain key data of most scientific papers in existence. GENtle includes a module to query PubMed, offering an improved interface compared to the online service. PubMed can be found at http://www.pubmed.gov.

4.4.3 Web-based tools

GENtle offers to run an amino acid or nucleotide sequence through a variety of online tools, including functions for analysis and prediction of primary structure, topology, motifs and functions, secondary structure, and posttranslational modification:

- Phobius and Poly-Phobius (transmembrane prediction) (http://phobius.cgb.ki.se)
- Motif scan (http://myhits.isb-sib.ch)
- P-val FPScan (http://umber.sbs.man.ac.uk)
- ELM (Functional site prediction) (http://elm.eu.org)
- Jpred (http://www.compbio.dundee.ac.uk)
- GOR and HNN (http://npsa-pbil.ibcp.fr)
- Phosphorylation states mW+pI (http://scansite.mit.edu)
- MitoProt II (http://ihg.gsf.de)
- Myristoylator (http://www.expasy.org)
- Sulfinator (http://www.expasy.org)
- SUMOplot (http://www.abgent.com)
- 2ZIP (leucine zipper prediction) (http://2zip.molgen.mpg.de)
- TargetP (subcellular localization) (http://www.cbs.dtu.dk)
- DGPI (http://129.194.185.165/dgpi)
- SAPS (http://www.isrec.isb-sib.ch)

- NEBcutter (http://tools.neb.com)
- Nomad (http://tools.neb.com)
- MultAlin (http://prodes.toulouse.inra.fr)
- WebLogo (http://weblogo.berkeley.edu)
- Translate (http://www.expasy.org)
- PrePS (Prenylation) (http://mendel.imp.ac.at)
- PlasMapper (http://wishart.biology.ualberta.ca)

4.4.4 SourceForge

SourceForge is a well-known online service that offers free hosting of open source projects. This includes source code versioning systems like CVS and SVN, management of released files, project homepage hosting, bug tracking systems, and discussion forums.

For GENtle, I have created a SourceForge project („gentle-m“) to ease synchronization of edits between my different development platforms, and to take advantage of the versioning via CVS, which also helps in distributing up-to-date source files. Also, this will allow for the collaboration of more programmers on GENtle development in the future. The project page is available at http://sourceforge.net/projects/gentle-m.

4.4.5 WikiBooks

WikiBooks is a spin-off from Wikipedia, a collaboratively edited encyclopedia under a free license (GNU Free Documentation License, GFDL). Similarly, WikiBooks aims to create free textbooks, including software manuals. For GENtle, I have switched from an integrated help system to an online manual at WikiBooks. Not only does this provide up-to-date help to user, but allows them to contribute and improve the manual. The manual is available at http://en.wikibooks.org/wiki/GENtle_manual.

5 Results and Discussion

5.1 Code and database layout

5.1.1 Code statistics

As of September 2006, GENtle consists of more than 48.000 lines of C++ code written by myself, as well as an additional 49.000 lines of foreign C/C++ code and header files from ClustalW, TinyXML, Ncoils, sqlite, UReadSeq, and IPC. This does not include the code for used libraries (MySQL, sqlite, wxWidgets and associated image libraries, as well as standard C/C++ libraries). My own code is spread across 75 source files (not counting headers). In total, there are about 170 C++ classes, including foreign code. Due to the amount of code and its inherent complexity, algorithms will be described in detail only where deemed appropriate. For implementation details, consult the code documentation (→4.2.8) or the code itself.

5.1.2 Databases

In order to avoid code duplication, MySQL, sqlite2 and sqlite3 all use the same SQL structure (Figure 2) and commands, as implemented by a single class. The actual database calls are done within private methods, effectively capsuling and opaquing database access from the rest of the code.

GENtle comes with two databases : A “blank” database containing nothing but a list of restriction enzymes and proteases, and a database containing common vectors for the automatic annotation function. New sqlite3 databases are created by

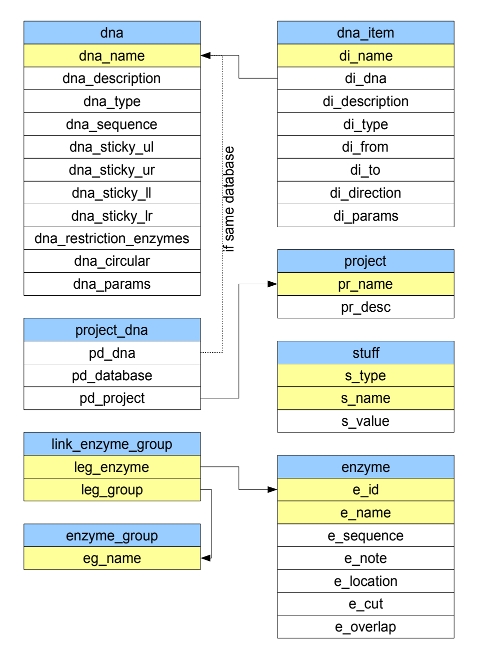
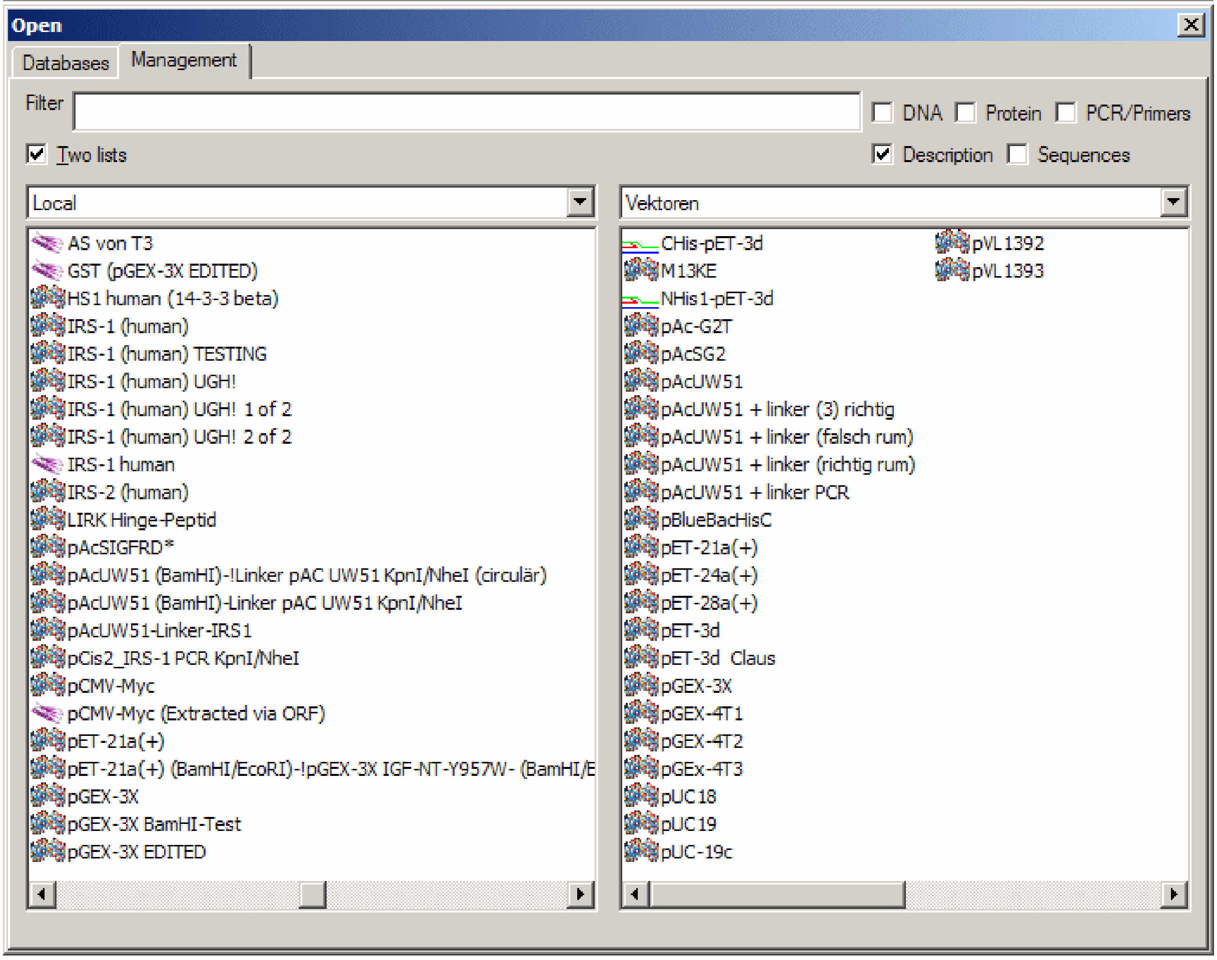


Figure 2: GENtle database schema

Table names are shown in blue, keys in yellow. Arrows indicate reference to a value in another table.

copying the blank database, while new MySQL databases are created through a set of SQL commands in a file also included with GENtle.



Open

Databases | Management

Filter ☐ DNA ☐ Protein ☐ PCR/Primers

☑ Two lists ☑ Description ☐ Sequences

Local

- AS von T3
- GST (pGEX-3X EDITED)
- HS1 human (14-3-3 beta)
- IRS-1 (human)
- IRS-1 (human) TESTING
- IRS-1 (human) UGH!
- IRS-1 (human) UGH! 1 of 2
- IRS-1 (human) UGH! 2 of 2
- IRS-1 human
- IRS-2 (human)
- LIRK Hinge-Peptid
- pAcSIGFRD*
- pAcUW51 (BamHI)-!Linker pAC UW51 KpnI/NheI (circulär)
- pAcUW51 (BamHI)-Linker pAC UW51 KpnI/NheI
- pAcUW51-Linker-IRS1
- pCis2_IRS-1 PCR KpnI/NheI
- pCMV-Myc
- pCMV-Myc (Extracted via ORF)
- pET-21a(+)
- pET-21a(+) (BamHI/EcoRI)-!pGEX-3X IGF-NT-Y957W- (BamHI/E
- pGEX-3X
- pGEX-3X BamHI-Test
- pGEX-3X EDITED

Vektoren

- CHis-pET-3d
- M13KE
- NHis1-pET-3d
- pAc-G2T
- pAcSG2
- pAcUW51
- pAcUW51 + linker (3) richtig
- pAcUW51 + linker (falsch rum)
- pAcUW51 + linker (richtig rum)
- pAcUW51 + linker PCR
- pBlueBacHisC
- pET-21a(+)
- pET-24a(+)
- pET-28a(+)
- pET-3d
- pET-3d Claus
- pGEX-3X
- pGEX-4T1
- pGEX-4T2
- pGEx-4T3
- pUC18
- pUC19
- pUC-19c
- pVL1392
- pVL1393

Figure 3: Database access.

Viewing two databases simultaneously (left and right, respectively), filter options on top.

Originally, SQLITE2 was used exclusively, however, this became a problem as SQLITE2 does not store data exceeding 1MB in size. Thus, SQLITE2 is included for backwards compatability; SQLITE2 databases are converted to SQLITE3 on-the-fly if there is an attempt to break the 1MB barrier by the user. New databases are created exclusively as SQLITE3 or MySQL.

All databases are accessed through a common interface (Figure 3). Data can be filtered by keywords and data types (DNA, amino acid sequences, primers, alignments). Keywords can be searched for in entry names, descriptions, and sequences. Data can be moved or copied from one database to another by drag-and-drop, renamed or deleted. As a safety function against accidentally overwriting an entry, saving an entry is blocked if there already exists an entry with a different sequence under that name in the database.

5.1.3 Sequence data

Internally, sequence data of both DNA and amino acids is stored in a single class, `TVector`, which contains all basic methods for annotating, analyzing and manipulating the respective sequence. While this at first seems to violate the meaning of the class inheritance model towards more specific class types, it proved to be quite useful in practice, since the class can act as a universal sequence storage.

5.1.4 Sequence display

A single class `SequenceCanvas` is used to display all sequences and associated information, such as restriction sites, features, and resulting amino acids. To that end, a sequence display consists of several lines of different types, with each type being represented by its own class. These line classes are all children of the class `SeqBasic`, which carries a rudimentary set of member variables and methods. Sequence lines are arranged each time one of them is changed, and otherwise displayed in their latest arrangement. Under certain conditions, the arrangement caches the position of the line elements, for example, in sequencer data view. Usually, though, element positions are calculated upon drawing, which can save huge amounts of memory, especially when dealing with larger sequences.

Sequence displays scroll vertically by default, simulating a page-like layout, but most of them can be turned into a horizontal scrolling mode. All sequence displays can be adapted in font size, and printed, copied, and saved as a bitmap image.

5.1.5 Algorithm runtime

Several of the algorithms in GENtle are used extensively, thus influencing overall performance. Runtime estimations [15] for these are as follows:

- Sequence lines display runtime is $\Omega(n)$ with n being the number of nucleotides or amino acids. An exception to this is the sequence line for features, which is worse, and thus deactivated by default when displaying very large sequences.
- Translation, likewise, is $\Omega(n)$ with n being the number of nucleotides in the sequence.
- Restriction site search is $\Omega(n \times m)$ with n being the number of nucleotides in the sequence, and m being the number of restriction endonucleases to check. However, due to several cutoff conditions, runtime is generally a lot better. Similar behaviour is observed for protease cleave site detection.

5.1.6 Updates

Updates of GENtle are automatically searched for online when GENtle starts, offering the user to download and run the update automatically. A brief comment highlights the changes, so the user can decide if this update is worth the time and bandwidth.

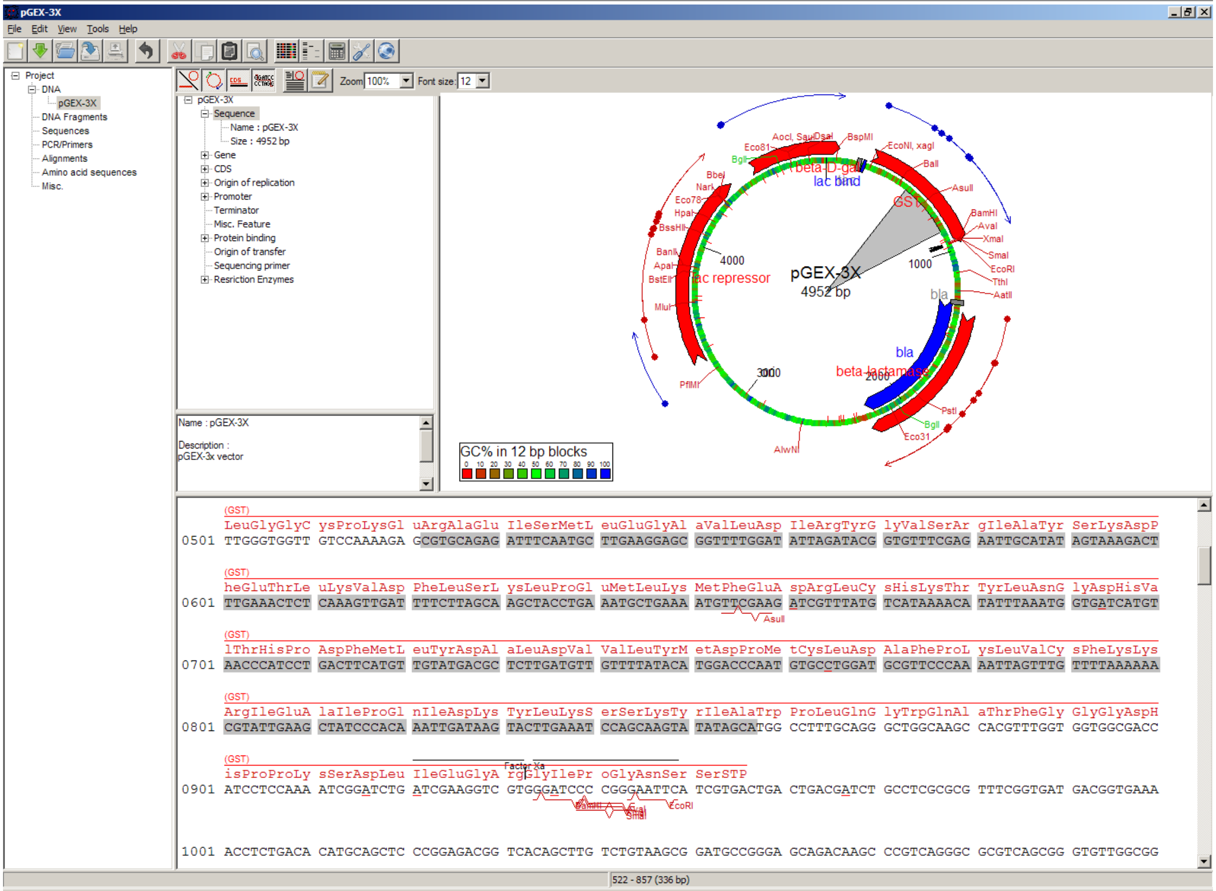
Internally, program and database versions are handled separately. Each database carries a version marked, specifying which database version an accessing GENtle instance must understand in order to operate the database without problems. An old version of GENtle is thus prevented from accidentally damaging a newer database schema. If such a condition occurs, the user is notified to update the GENtle version, and the GENtle instance can access the database for reading only.

5.2 Modules

GENtle is divided into modules, offering views and functions suitable for that specific context. While many GENtle modules are sequence-based, there are several modules based on images or spreadsheet data. Each open module is listed in a categorized tree structure always visible at the left side of the GENtle application, and is internally represented by its respective class. All module classes are based on the parent class `ChildBase`, which offers basic methods and member variables, including a common internal interface.

5.2.1 *DNA sequences*

One of the core modules of GENtle is the DNA sequence module (Figure 4). It allows for display, analysis, and manipulation of a DNA sequence.



Name : pGEX-3X
Size : 4952 bp

Name : pGEX-3X
Description :
pGEX-3x vector

```
     (GST)
     LeuGlyGlyC ysProLysGl uArgAlaGlu IleSerMetL euGluGlyAl aValLeuAsp IleArgTyrG lyValSerAr gIleAlaTyr SerLysAspP
0501 TTGGGTGGTT GTCCAAAAGA GCGTGCAGAG ATTTCAATGC TTGAAGGAGC GGTTTTGGAT ATTAGATACG GTGTTTCGAG AATTGCATAT AGTAAAGACT

     (GST)
     heGluThrLe uLysValAsp PheLeuSerL ysLeuProGl uMetLeuLys MetPheGluA spArgLeuCy sHisLysThr TyrLeuAsnG lyAspHisVa
0601 TTGAAACTCT CAAAGTTGAT TTTCTTAGCA AGCTACCTGA AATGCTGAAA ATGTTCGAAG ATCGTTTATG TCATAAAACA TATTTAAATG GTGATCATGT
                                                                AsuII
     (GST)
     lThrHisPro AspPheMetL euTyrAspAl aLeuAspVal ValLeuTyrM etAspProMe tCysLeuAsp AlaPheProL ysLeuValCy sPheLysLys
0701 AACCCATCCT GACTTCATGT TGTATGACGC TCTTGATGTT GTTTTATACA TGGACCCAAT GTGCCTGGAT GCGTTCCCAA AATTAGTTTG TTTTAAAAAA

     (GST)
     ArgIleGluA laIleProGl nIleAspLys TyrLeuLysS erSerLysTy rIleAlaTrp ProLeuGlnG lyTrpGlnAl aThrPheGly GlyGlyAspH
0801 CGTATTGAAG CTATCCCACA AATTGATAAG TACTTGAAAT CCAGCAAGTA TATAGCATGG CCTTTGCAGG GCTGGCAAGC CACGTTTGGT GGTGGCGACC

     (GST)                          Factor Xa
     isProProLy sSerAspLeu IleGluGlyA rgGlyIlePr oGlyAsnSer SerSTP
0901 ATCCTCCAAA ATCGGATCTG ATCGAAGGTC GTGGGATCCC CGGGAATTCA TCGTGACTGA CTGACGATCT GCCTCGCGCG TTTCGGTGAT GACGGTGAAA
                                          BamHI   AvaI  EcoRI
                                                  SmaI

1001 ACCTCTGACA CATGCAGCTC CCGGAGACGG TCACAGCTTG TCTGTAAGCG GATGCCGGGA GCAGACAAGC CCGTCAGGGC GCGTCAGCGG GTGTTGGCGG
```

522 - 857 (336 bp)

Figure 4: The DNA module.

The module tree is located on the very left, the DNA/plasmid map on the upper right, the DNA properties tree on the upper left, the DNA sequence on the bottom. A marked part of the sequence is shown in gray in both map and sequence. The map shows GC contents, methylation sites, and open reading frames. Both map and sequence show sites of restriction enzymes and annotated features. The sequence shows a Factor Xa protease site.

The sequence display shows the DNA sequence itself, and optionally the complementary 3'→5' DNA sequence, annotated features, *E. coli* methylation sites [16], restriction endonuclease sites, the amino acid sequence resulting from the DNA in either manual or automatic (feature-based) reading frame(s) including potential protease sites.

The map display above the sequence shows the same but for protease sites, and additionally can display open reading frames and GC contents, as well as sticky ends of linear sequences, if applicable. The map can be printed, saved, and copied to the clipboard as an image.

DNA can be marked in both displays, and edited directly in the sequence display. All information depending on the DNA changes instantaneously with the edited DNA. Thus, the resulting amino acid sequence changes with an altered DNA sequence. If the introduced change introduces or removes a protease cleavage site for a selected protease, that site will appear or disappear, respectively, in the sequence display. Likewise, newly introduced or removed restriction sites will show up or disappear accordingly.

The sequences can be edited, copied or cut manually, transformed (inverse and/or complementary sequence), and strands can be extracted. A search function finds DNA stretches on either strand, restriction enzyme sites, amino acid stretches in all reading frames, and annotated features. Results can be temporarily highlighted in the sequence.

Amino acid sequences can be extracted via features or manually set reading frame and DNA sequence marking. The modul can mark potential siRNA duplexes (→4.3.2.2), automatically annotate DNA sequences from both an included set of standard vectors and own sequences in a database, and anneal primers from a database which could function as sequencing primers to the sequence. BLAST searches (→4.4.1) can be performed for DNA or amino acid sequences.

Additional information about the DNA sequence, such as name and size, a tree structure with features, a list of restriction enzymes, and a description are shown left of the map display.

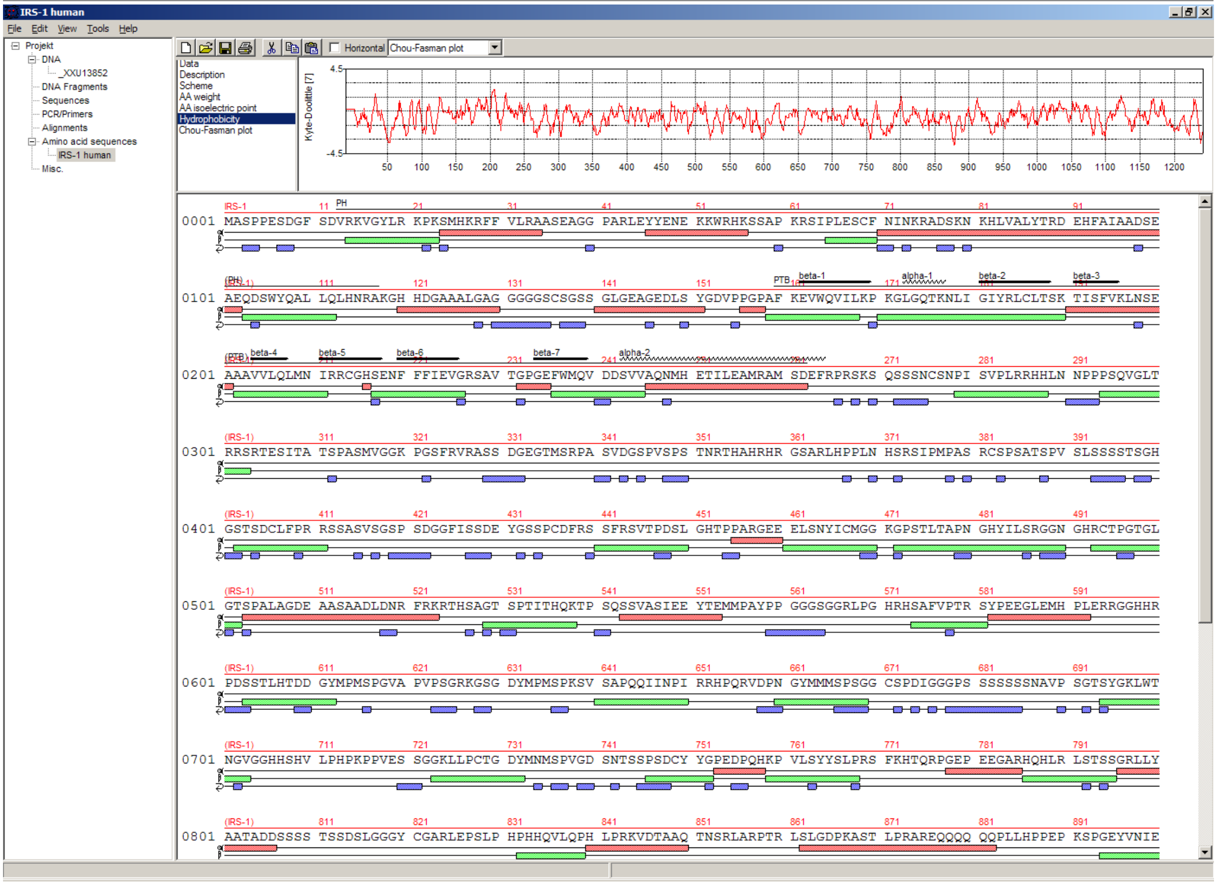
Like in most modules, a multitude of functions can be applied to objects such as annotated features, restriction sites, and the (marked) sequence through a context menu, which will appear when clicking on such an object with the right mouse button. This works for all object types in the tree and map; the context menu of the sequence display is limited to sequence operations, as features and restriction sites are sometimes hard to discern with the mouse cursor.

The DNA sequence module also supports the middle (“wheel”) mouse button, where available, to invoke the Restriction Assistant on restriction sites and “mark and show” the sequence of annotated features.

5.2.2 Amino acid sequences

The amino acid sequence module is, in structure, similar to the DNA sequence module (Figure 5). The sequence and its annotated features are the core of the sequence display. The tree and map have been replaced by a multiple function selector, offering a map-like schema of the sequence, a page with notes, an automatically generated set of key data of the sequence, and several functional plots of sequence properties. These include amino acid weight, isoelectric point, hydrophobicity (→4.3.2.6), Chou-Fasman secondary structure prediction (→4.3.2.5), and coiled-coil prediction (→4.3.2.1). Compressed versions of the latter two, as well as the molecular amino acid structures, can also be shown “in line”, directly below the sequence.

The data sets and calculation methods for the key data calculation were taken from ExPASy [17]. These include number of amino acids, molecular weight (mW), estimated isoelectic point (pI), number of positively and negatively charged amino acids, detailed count of composing atoms and amino acids, estimated half-life in different organisms, and the grand average of hydropathicity (GRAVY).



```
0001 MASPPESDGF SDVRKVGYLR KPKSMHKRFF VLRAASEAGG PARLEYYENE KKWRHKSSAP KRSIPLESCF NINKRADSKN KHLVALYTRD EHFAIAADSE
0101 AEQDSWYQAL LQLHNRAKGH HDGAAALGAG GGGGSCSGSS GLGEAGEDLS YGDVPPGPAF KEVWQVILKP KGLGQTKNLI GIYRLCLTSK TISFVKLNSE
0201 AAAVVLQLMN IRRCGHSENF FFIEVGRSAV TGPGEFWMQV DDSVVAQNMH ETILEAMRAM SDEFRPRSKS QSSSNCSNPI SVPLRRHHLN NPPPSQVGLT
0301 RRSRTESITA TSPASMVGGK PGSFRVRASS DGEGTMSRPA SVDGSPVSPS TNRTHAHRHR GSARLHPPLN HSRSIPMPAS RCSPSATSPV SLSSSSTSGH
0401 GSTSDCLFPR RSSASVSGSP SDGGFISSDE YGSSPCDFRS SFRSVTPDSL GHTPPARGEE ELSNYICMGG KGPSTLTAPN GHYILSRGGN GHRCTPGTGL
0501 GTSPALAGDE AASAADLDNR FRKRTHSAGT SPTITHQKTP SQSSVASIEE YTEMMPAYPP GGGSGGRLPG HRHSAFVPTR SYPEEGLEMH PLERRGGHHR
0601 PDSSTLHTDD GYMPMSPGVA PVPSGRKGSG DYMPMSPKSV SAPQQIINPI RRHPQRVDPN GYMMMSPSGG CSPDIGGGPS SSSSSSNAVP SGTSYGKLWT
0701 NGVGGHHSHV LPHPKPPVES SGGKLLPCTG DYMNMSPVGD SNTSSPSDCY YGPEDPQHKP VLSYYSLPRS FKHTQRPGEP EEGARHQHLR LSTSSGRLLY
0801 AATADDSSSS TSSDSLGGGY CGARLEPSLP HPHHQVLQPH LPRKVDTAAQ TNSRLARPTR LSLGDPKAST LPRAREQQQQ QQPLLHPPEP KSPGEYVNIE
```

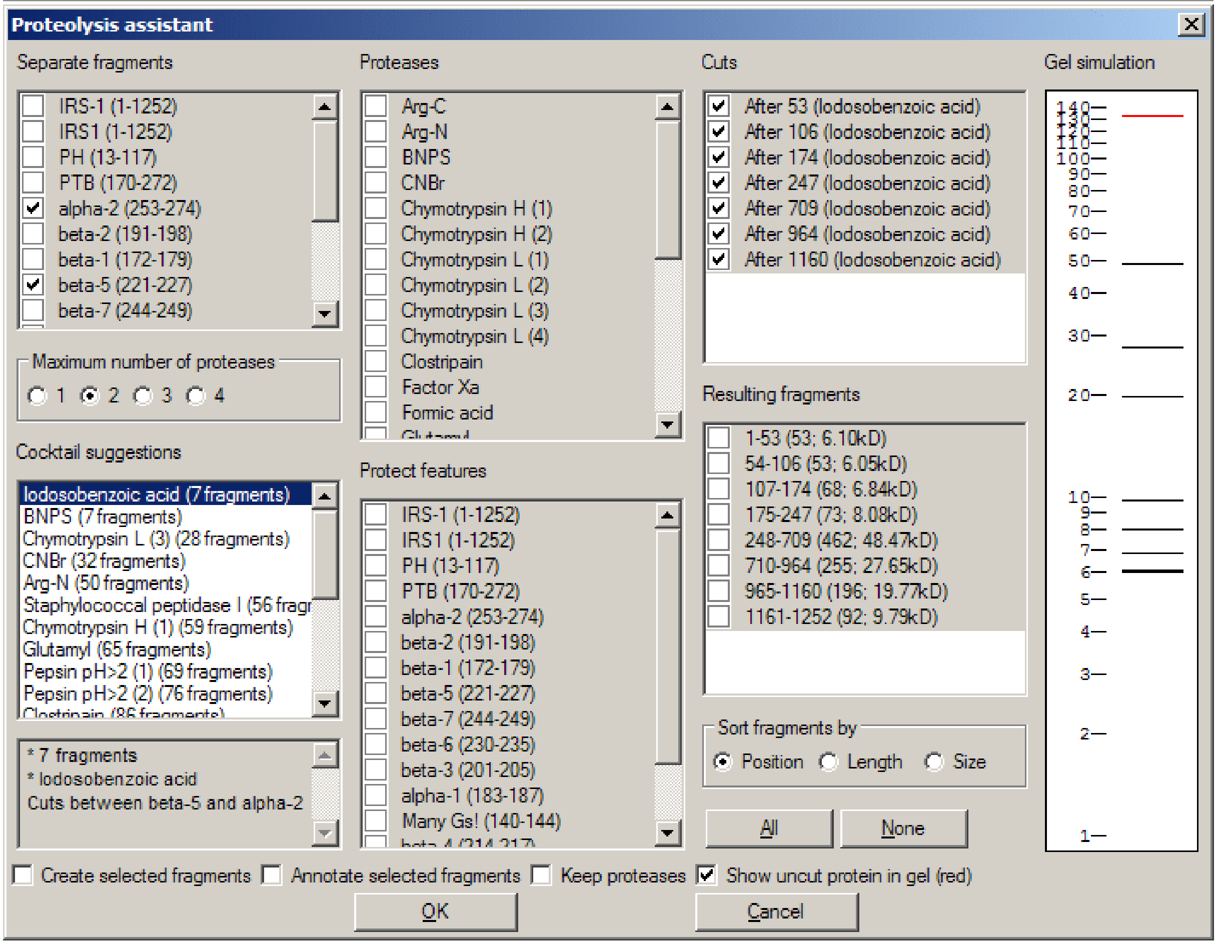
Figure 5: The amino acid module.

The major part of the module shows the amino acid sequence itself. Automatically numbered annotated features are shown above the sequence. Below the sequence, a simplified Chou-Fasman-plot [11] shows predicted α helices (red), β sheets (green), and turns (blue). Above the sequence display, the function selector is set to hydrophobicity, showing the respective plot in red.

Many actions, such as editing and searching the sequence, work similarly to the DNA sequence module as well, with obvious adaptions (e.g., no searching in different reading frames). Likewise, update of features, key data, and protease sites is done instantaneously as well.

Among the functions to perform on amino acid sequences are BLAST searches (→4.4.1), calculations for photometric analysis of concentration and purity, “backtranslating” into DNA using standard or species-specific codon preferences, an isotopic pattern calculator (→4.3.2.3) to predict mass spectrometer data for short peptides, and a Proteolysis Assistant.

5.2.2.1 *Proteolysis assistant*



Separate fragments
- □ IRS-1 (1-1252)
- □ IRS1 (1-1252)
- □ PH (13-117)
- □ PTB (170-272)
- ☑ alpha-2 (253-274)
- □ beta-2 (191-198)
- □ beta-1 (172-179)
- ☑ beta-5 (221-227)
- □ beta-7 (244-249)

Maximum number of proteases
○ 1 ● 2 ○ 3 ○ 4

Cocktail suggestions
- Iodosobenzoic acid (7 fragments)
- BNPS (7 fragments)
- Chymotrypsin L (3) (28 fragments)
- CNBr (32 fragments)
- Arg-N (50 fragments)
- Staphylococcal peptidase I (56 fragr
- Chymotrypsin H (1) (59 fragments)
- Glutamyl (65 fragments)
- Pepsin pH>2 (1) (69 fragments)
- Pepsin pH>2 (2) (76 fragments)
- Clostripain (86 fragments)

* 7 fragments
* Iodosobenzoic acid
Cuts between beta-5 and alpha-2

Proteases
- □ Arg-C
- □ Arg-N
- □ BNPS
- □ CNBr
- □ Chymotrypsin H (1)
- □ Chymotrypsin H (2)
- □ Chymotrypsin L (1)
- □ Chymotrypsin L (2)
- □ Chymotrypsin L (3)
- □ Chymotrypsin L (4)
- □ Clostripain
- □ Factor Xa
- □ Formic acid
- □ Glutamyl

Protect features
- □ IRS-1 (1-1252)
- □ IRS1 (1-1252)
- □ PH (13-117)
- □ PTB (170-272)
- □ alpha-2 (253-274)
- □ beta-2 (191-198)
- □ beta-1 (172-179)
- □ beta-5 (221-227)
- □ beta-7 (244-249)
- □ beta-6 (230-235)
- □ beta-3 (201-205)
- □ alpha-1 (183-187)
- □ Many Gs! (140-144)
- □ beta-4 (214-217)

Cuts
- ☑ After 53 (Iodosobenzoic acid)
- ☑ After 106 (Iodosobenzoic acid)
- ☑ After 174 (Iodosobenzoic acid)
- ☑ After 247 (Iodosobenzoic acid)
- ☑ After 709 (Iodosobenzoic acid)
- ☑ After 964 (Iodosobenzoic acid)
- ☑ After 1160 (Iodosobenzoic acid)

Resulting fragments
- □ 1-53 (53; 6.10kD)
- □ 54-106 (53; 6.05kD)
- □ 107-174 (68; 6.84kD)
- □ 175-247 (73; 8.08kD)
- □ 248-709 (462; 48.47kD)
- □ 710-964 (255; 27.65kD)
- □ 965-1160 (196; 19.77kD)
- □ 1161-1252 (92; 9.79kD)

Sort fragments by
● Position ○ Length ○ Size

□ Create selected fragments □ Annotate selected fragments □ Keep proteases ☑ Show uncut protein in gel (red)

Figure 6: The Proteolysis Assistant.

Features to separate and list with suggestions (left column), active proteases and protected features (middle column), cleavage sites and resulting fragments (right column), and virtual gel (far right).

Protease cleavage sites can be displayed directly in the amino acid sequence, however, that requires the user to chose a set of proteases. The Proteolysis Assistant does the opposite, allowing the user to find suitable proteases by simulating protease cleavage *en masse*. Based on which annotated features the user wants to separate, the assistant can suggest one or a combination of enzymes to fit the task as closely as possible. For such a suggestion, or manually chosen proteases, the resulting protein fragments are calculated. Features can be marked as proteolytically stable, influencing the resulting fragments. Likewise, individual cleavage sites can be turned off. The resulting fragments are displayed in both a list and a small virtual gel. Selected fragments can be annotated as such in the original sequence or used for new amino acid sequence modules.

5.2.3 Primer design and virtual PCR

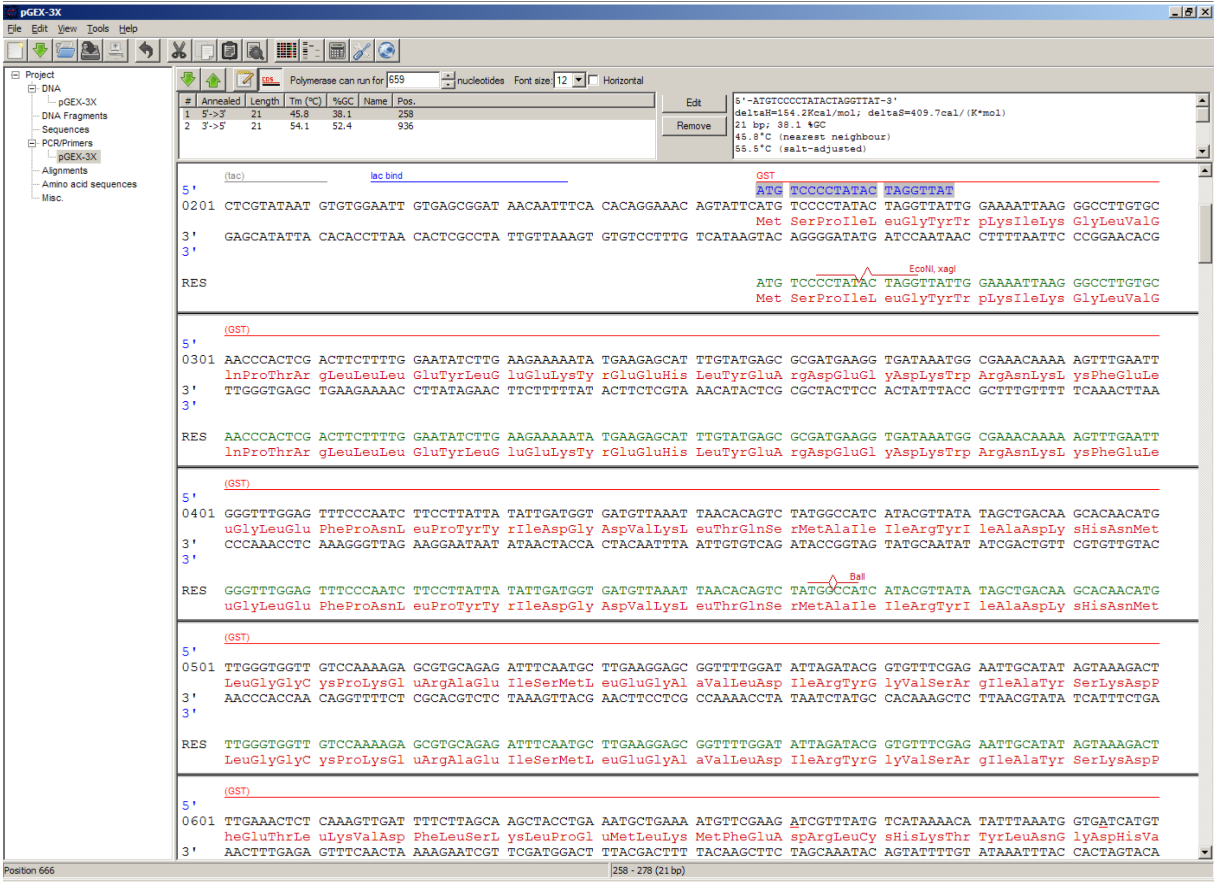
GENtle features a module for primer design, optimization, and virtual PCR (Figure 7). Polymerase chain reaction [18] has become an essential tool in most of the genetics-related lab work. The primer design/virtual PCR module can be invoked from the DNA module, and suggest none, one, or two primers to limit the part of the sequence to amplify. Also, when a stretch of three nucleotides (a codon) is selected, mutagenesis primers can be suggested automatically.

5.2.3.1 Manual primer design

Similar to the amino acid module, the PCR module is dominated by the sequence display, which contains the template DNA in both 5'→3' and 3'→5' notation, annotated features, the primers next to the DNA, the resulting DNA sequence (the PCR product) with restriction enzyme sites, as well as the resulting amino acid sequence(s) for both the template and the product DNA.

The primers can be edited directly in the sequence display; all other sequence information is protected (template DNA, features, and amino acids) or recalculated depending on template and primer DNA (PCR product, resulting amino acids, restriction sites). Primers can, besides through automatic suggestions, be imported or added manually. Primers can be saved in the database.

Both the annealing function after import (to place the primers where they will most likely anneal) and the melting temperature calculations are based on the 3' end of the primer that matches the template sequence, as this is the part most important for the polymerase and, thus, the PCR product generation. As a consequence, every nucleotide 5' of the first mutation (seen from the 3' end) is ignored for these processes, which may lead to unusually low melting temperature calculations. An improved algorithm is being worked on.



Polymerase can run for 659 nucleotides Font size: 12 □ Horizontal

#	Annealed	Length	Tm (°C)	%GC	Name	Pos.
1	5'->3'	21	45.8	38.1		258
2	3'->5'	21	54.1	52.4		936

```
5'-ATGTCCCCTATACTAGGTTAT-3'
deltaH=154.2Kcal/mol; deltaS=409.7cal/(K*mol)
21 bp; 38.1 %GC
45.8°C (nearest neighbour)
55.5°C (salt-adjusted)
```

```
     (tac)            lac bind                                          GST
5'                                                                  ATG TCCCCTATAC TAGGTTAT
0201 CTCGTATAAT GTGTGGAATT GTGAGCGGAT AACAATTTCA CACAGGAAAC AGTATTCATG TCCCCTATAC TAGGTTATTG GAAAATTAAG GGCCTTGTGC
                                                                    Met SerProIleL euGlyTyrTr pLysIleLys GlyLeuValG
3'   GAGCATATTA CACACCTTAA CACTCGCCTA TTGTTAAAGT GTGTCCTTTG TCATAAGTAC AGGGGATATG ATCCAATAAC CTTTTAATTC CCGGAACACG
3'

                                                                                EcoNI, xagI
RES                                                                 ATG TCCCCTATAC TAGGTTATTG GAAAATTAAG GGCCTTGTGC
                                                                    Met SerProIleL euGlyTyrTr pLysIleLys GlyLeuValG

     (GST)
5'
0301 AACCCACTCG ACTTCTTTTG GAATATCTTG AAGAAAAATA TGAAGAGCAT TTGTATGAGC GCGATGAAGG TGATAAATGG CGAAACAAAA AGTTTGAATT
     lnProThrAr gLeuLeuLeu GluTyrLeuG luGluLysTy rGluGluHis LeuTyrGluA rgAspGluGl yAspLysTrp ArgAsnLysL ysPheGluLe
3'   TTGGGTGAGC TGAAGAAAAC CTTATAGAAC TTCTTTTTAT ACTTCTCGTA AACATACTCG CGCTACTTCC ACTATTTACC GCTTTGTTTT TCAAACTTAA
3'

RES  AACCCACTCG ACTTCTTTTG GAATATCTTG AAGAAAAATA TGAAGAGCAT TTGTATGAGC GCGATGAAGG TGATAAATGG CGAAACAAAA AGTTTGAATT
     lnProThrAr gLeuLeuLeu GluTyrLeuG luGluLysTy rGluGluHis LeuTyrGluA rgAspGluGl yAspLysTrp ArgAsnLysL ysPheGluLe

     (GST)
5'
0401 GGGTTTGGAG TTTCCCAATC TTCCTTATTA TATTGATGGT GATGTTAAAT TAACACAGTC TATGGCCATC ATACGTTATA TAGCTGACAA GCACAACATG
     uGlyLeuGlu PheProAsnL euProTyrTy rIleAspGly AspValLysL euThrGlnSe rMetAlaIle IleArgTyrI leAlaAspLy sHisAsnMet
3'   CCCAAACCTC AAAGGGTTAG AAGGAATAAT ATAACTACCA CTACAATTTA ATTGTGTCAG ATACCGGTAG TATGCAATAT ATCGACTGTT CGTGTTGTAC
3'
                                                                         BalI
RES  GGGTTTGGAG TTTCCCAATC TTCCTTATTA TATTGATGGT GATGTTAAAT TAACACAGTC TATGGCCATC ATACGTTATA TAGCTGACAA GCACAACATG
     uGlyLeuGlu PheProAsnL euProTyrTy rIleAspGly AspValLysL euThrGlnSe rMetAlaIle IleArgTyrI leAlaAspLy sHisAsnMet

     (GST)
5'
0501 TTGGGTGGTT GTCCAAAAGA GCGTGCAGAG ATTTCAATGC TTGAAGGAGC GGTTTTGGAT ATTAGATACG GTGTTTCGAG AATTGCATAT AGTAAAGACT
     LeuGlyGlyC ysProLysGl uArgAlaGlu IleSerMetL euGluGlyAl aValLeuAsp IleArgTyrG lyValSerAr gIleAlaTyr SerLysAspP
3'   AACCCACCAA CAGGTTTTCT CGCACGTCTC TAAAGTTACG AACTTCCTCG CCAAAACCTA TAATCTATGC CACAAAGCTC TTAACGTATA TCATTTCTGA
3'

RES  TTGGGTGGTT GTCCAAAAGA GCGTGCAGAG ATTTCAATGC TTGAAGGAGC GGTTTTGGAT ATTAGATACG GTGTTTCGAG AATTGCATAT AGTAAAGACT
     LeuGlyGlyC ysProLysGl uArgAlaGlu IleSerMetL euGluGlyAl aValLeuAsp IleArgTyrG lyValSerAr gIleAlaTyr SerLysAspP

     (GST)
5'
0601 TTGAAACTCT CAAAGTTGAT TTTCTTAGCA AGCTACCTGA AATGCTGAAA ATGTTCGAAG ATCGTTTATG TCATAAAACA TATTTAAATG GTGATCATGT
     heGluThrLe uLysValAsp PheLeuSerL ysLeuProGl uMetLeuLys MetPheGluA spArgLeuCy sHisLysThr TyrLeuAsnG lyAspHisVa
3'   AACTTTGAGA GTTTCAACTA AAAGAATCGT TCGATGGACT TTACGACTTT TACAAGCTTC TAGCAAATAC AGTATTTTGT ATAAATTTAC CACTAGTACA
```

Figure 7: The primer design and virtual PCR module.

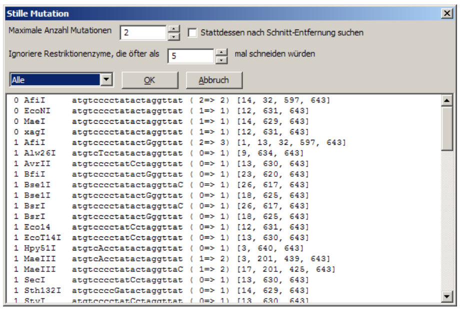
The large sequence display shows annotated features, 5' primer (blue, marked on gray background), 5'→3' template DNA (black), amino acid sequence of the template (red), 3'→5' template DNA (black), 3' primer (blue, not visible in this figure), resulting PCR product DNA (green) with restriction sites (above, red), and resulting amino acid sequence (red, below). At the top of the window, the primer list is located on the left, and the key data display of the primer selected in that list is located on the right.

To circumvent the variations of different polymerases and elongation times, I added a parameter for elongation length, which is suggested automatically depending on the primers and their annealing positions, but can be changed manually.

A list of primers is shown above the sequence display, and for each primer, key data can be calculated and displayed. Like the PCR product and associated information, this data is updated instantaneously when editing a primer, and includes melting temperatures calculated by three different algorithms (GC method, salt-adjusted, and nearest neighbour [19]), primer length, GC contents in percent, and the most likely self-annealing primer dimer.

5.2.3.2 *Silent mutagenesis*

Silent mutagenesis is a standard method to validate the success of a PCR. A primer is altered in a way that introduces a new restriction site within the PCR product without changing its amino acid sequence.


Stille Mutation
Maximale Anzahl Mutationen
Stattdessen nach Schnitt-Entfernung suchen
Ignoriere Restriktionenzyme, die öfter als
mal schneiden würden
OK
Abbruch

Maximale Anzahl Mutationen: 2
□ Stattdessen nach Schnitt-Entfernung suchen
Ignoriere Restriktionenzyme, die öfter als 5 mal schneiden würden
Alle

```
 0 AfiI     atgtccccatatactaggttat ( 2=> 2) [14, 32, 597, 643]
 0 EcoNI    atgtccccatatactaggttat ( 1=> 1) [12, 631, 643]
 0 MaeI     atgtccccatatactaggttat ( 1=> 1) [14, 629, 643]
 0 xagI     atgtccccatatactaggttat ( 1=> 1) [12, 631, 643]
 1 AfiI     atgtccccatatactGggttat ( 2=> 3) [1, 13, 32, 597, 643]
 1 Alw26I   atgtcTcctatactaggttat ( 0=> 1) [9, 634, 643]
 1 AvrII    atgtccccatatCctaggttat ( 0=> 1) [13, 630, 643]
 1 BfiI     atgtccccatatactGggttat ( 0=> 1) [23, 620, 643]
 1 Bse1I    atgtccccatatactaggttaC ( 0=> 1) [26, 617, 643]
 1 Bse1I    atgtccccatatactGggttat ( 0=> 1) [18, 625, 643]
 1 BsrI     atgtccccatatactaggttaC ( 0=> 1) [26, 617, 643]
 1 BsrI     atgtccccatatactGggttat ( 0=> 1) [18, 625, 643]
 1 Eco14    atgtccccatatCctaggttat ( 0=> 1) [12, 631, 643]
 1 EcoT14I  atgtccccatatCctaggttat ( 0=> 1) [13, 630, 643]
 1 Hpy51I   atgtcAcctatactaggttat ( 0=> 1) [3, 640, 643]
 1 MaeIII   atgtcAcctatactaggttat ( 1=> 2) [3, 201, 439, 643]
 1 MaeIII   atgtccccatatactaggttaC ( 1=> 2) [17, 201, 425, 643]
 1 SecI     atgtccccatatCctaggttat ( 0=> 1) [13, 630, 643]
 1 Sth132I  atgtccccGatactaggttat ( 0=> 1) [14, 629, 643]
 1 StyI     atgtccccatatCctaggttat ( 0=> 1) [13, 630, 643]
```

Figure 8: Silent Mutagenesis Assistant.

Settings on top, list of potential silent mutageneses (bottom).

A special dialog (the Silent Mutagenesis Assistant, Figure 8) allows to find suitable mutations for a primer. It searches for silent mutations parameterized by enzyme set, total number of cuts, and required number of base exchanges. Alternatively, it can also search for removal of a restriction site instead.

A list suggests the suitable mutations, showing the endonucleases concerned, the needed base exchanges, the numbers of cuts, both before and after the mutation, the altered sequence of the primer, as well as the number and size of the DNA fragments after restriction with the respective endonuclease. The user can select one of the suggestions, which inserts the selected mutations into the primer, and forces the display of the new restriction site.

5.2.3.3 *Primer optimization*

To improve primers within a parameter range, all primer variations for 3' and 5' variation, minimum and maximum values for both primer length and melting temperature can be calculated, weighted, and displayed. The user can preview key data of any of the new primers, and select one of them, updating the primer in the sequence display accordingly.

5.2.3.4 Troubleshooting assistant

Primers might exhibit a range of problems during the actual PCR process. To prevent problems, or to diagnose existing ones, a PCR Troubleshooting Assistant can check for optimal GC contents, self- and cross-dimers, stability, specificity, repeats, and GC ends. Most of the algorithms are based on free energy calculations based on the data and formulas of SantaLucia [19].

Some of the used algorithms always return a result, and some can issue a warning instead of a significant error. Where the values exceed predefined borders that indicate a major fault with a primer or the primer combination, the warning is emphasized by all-uppercase letters.

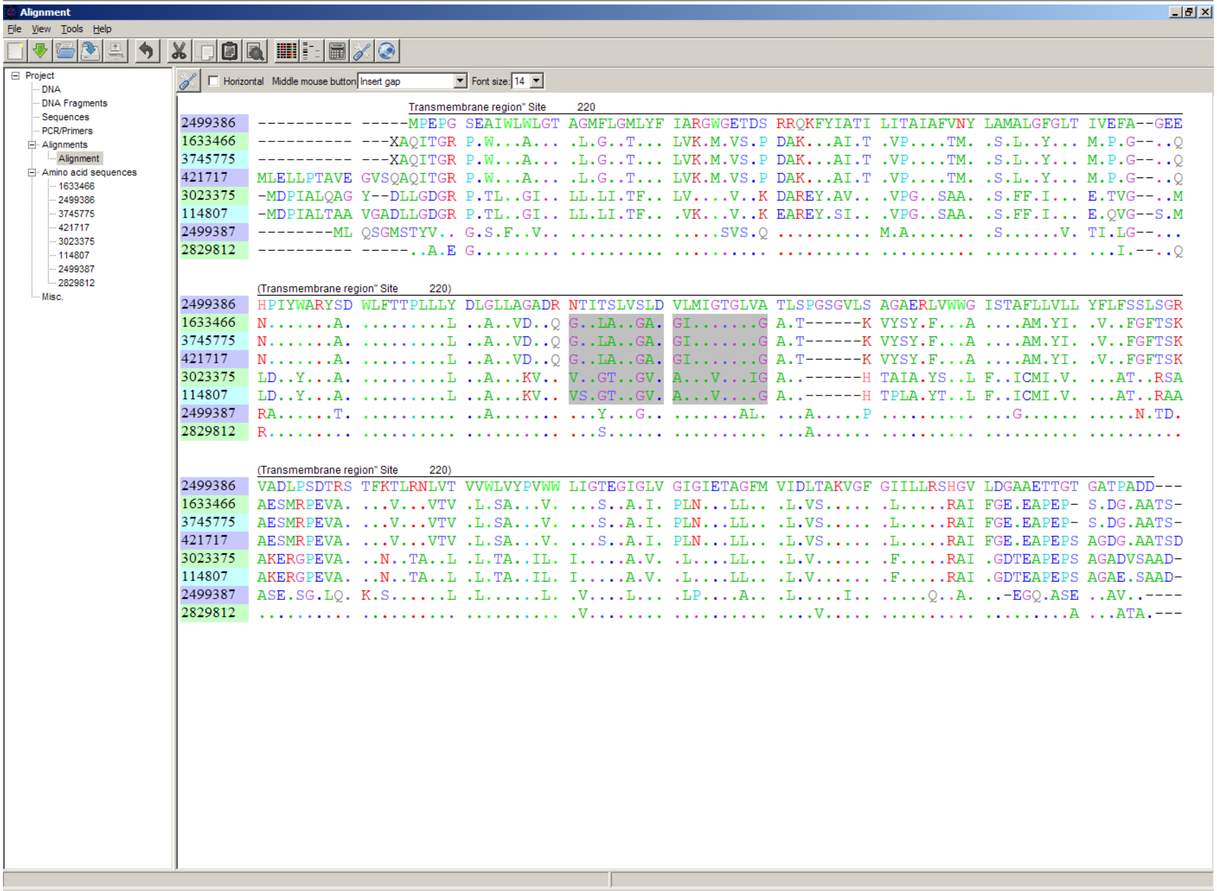
5.2.3.5 Virtual PCR

Once the primers are sufficiently optimized, the DNA sequence resulting from the PCR process (or its amino acid sequence, depending on the reading frame) can be calculated and used as a new sequence. All annotated features of the template DNA will be preserved where appropriate.

5.2.4 Alignments

Alignments (Figure 9) are essential for comparing sequences amongst each other. In GENtle, alignments can be generated anew from two or more sequences using one of several algorithms (→4.3.1), or already generated alignments can be opened from the database or imported in one of several file formats.

An alignment can be optimized manually by inserting and removing gaps. The order of sequences can be changed, and each sequence can display its original features. Display style can be altered through a combination of predefined color charts and manual editing of text, background, and border color and style, to better emphasize important parts of the alignment.



```
                          Transmembrane region" Site     220
2499386  ---------- -----MPEPG SEAIWLWLGT AGMFLGMLYF IARGWGETDS RRQKFYIATI LITAIAFVNY LAMALGFGLT IVEFA--GEE
1633466  ---------- ---XAQITGR P.W...A... .L.G..T... LVK.M.VS.P DAK...AI.T .VP....TM. .S.L..Y... M.P.G--..Q
3745775  ---------- ---XAQITGR P.W...A... .L.G..T... LVK.M.VS.P DAK...AI.T .VP....TM. .S.L..Y... M.P.G--..Q
421717   MLELLPTAVE GVSQAQITGR P.W...A... .L.G..T... LVK.M.VS.P DAK...AI.T .VP....TM. .S.L..Y... M.P.G--..Q
3023375  -MDPIALQAG Y--DLLGDGR P.TL..GI.. LL.LI.TF.. LV....V..K DAREY.AV.. .VPG..SAA. .S.FF.I... E.TVG--..M
114807   -MDPIALTAA VGADLLGDGR P.TL..GI.. LL.LI.TF.. .VK...V..K EAREY.SI.. .VPG..SAA. .S.FF.I... E.QVG--S.M
2499387  --------ML QSGMSTYV.. G.S.F..V.. .......... .....SVS.Q .......... M.A....... .S......V. TI.LG--...
2829812  ---------- -----..A.E G......... .......... .......... .......... .......... .......... ...I.--..Q

         (Transmembrane region" Site     220)
2499386  HPIYWARYSD WLFTTPLLLY DLGLLAGADR NTITSLVSLD VLMIGTGLVA TLSPGSGVLS AGAERLVWWG ISTAFLLVLL YFLFSSLSGR
1633466  N.......A. .........L ..A..VD..Q G..LA..GA. GI.......G A.T------K VYSY.F...A ....AM.YI. .V..FGFTSK
3745775  N.......A. .........L ..A..VD..Q G..LA..GA. GI.......G A.T------K VYSY.F...A ....AM.YI. .V..FGFTSK
421717   N.......A. .........L ..A..VD..Q G..LA..GA. GI.......G A.T------K VYSY.F...A ....AM.YI. .V..FGFTSK
3023375  LD..Y...A. .........L ..A...KV.. V..GT..GV. A...V...IG A..------H TAIA.YS..L F..ICMI.V. ...AT..RSA
114807   LD..Y...A. .........L ..A...KV.. VS.GT..GV. A...V....G A..------H TPLA.YT..L F..ICMI.V. ...AT..RAA
2499387  RA......T. .......... ..A....... ...Y...G.. .......AL. ...A.....P .......... ...G...... .....N.TD.
2829812  R......... .......... .......... ...S...... .......... ...A...... .......... .......... ..........

         (Transmembrane region" Site     220)
2499386  VADLPSDTRS TFKTLRNLVT VVWLVYPVWW LIGTEGIGLV GIGIETAGFM VIDLTAKVGF GIILLRSHGV LDGAAETTGT GATPADD---
1633466  AESMRPEVA. ...V...VTV .L.SA...V. ...S..A.I. PLN...LL.. .L.VS..... .L.....RAI FGE.EAPEP- S.DG.AATS-
3745775  AESMRPEVA. ...V...VTV .L.SA...V. ...S..A.I. PLN...LL.. .L.VS..... .L.....RAI FGE.EAPEP- S.DG.AATS-
421717   AESMRPEVA. ...V...VTV .L.SA...V. ...S..A.I. PLN...LL.. .L.VS..... .L.....RAI FGE.EAPEPS AGDG.AATSD
3023375  AKERGPEVA. ..N..TA..L .L.TA..IL. I.....A.V. .L....LL.. .L.V...... .F.....RAI .GDTEAPEPS AGADVSAAD-
114807   AKERGPEVA. ..N..TA..L .L.TA..IL. I.....A.V. .L....LL.. .L.V...... .F.....RAI .GDTEAPEPS AGAE.SAAD-
2499387  ASE.SG.LQ. K.S......L .L......L. .V....L... .LP....A.. .L.....I.. .....Q..A. ..-EGQ.ASE ..AV..----
2829812  .......... .......... .......... .V........ .......... ....V..... .......... .........A ...ATA.---
```

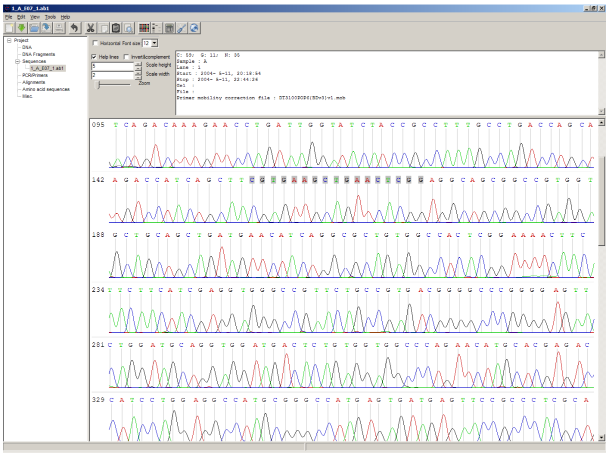
Figure 9: The Alignment module.

The sequence display shows aligned amino acid sequences. Amino acids identical to the one in the first row are shown as a dot. The first row shows annotated features. A sequence block is marked in gray.

5.2.4.1 Phylip

From an alignment can be created a phylogenetic tree, based on the similarity of two respective sequences to each other. GENtle can display phylogenetic trees generated by the PHYLIP [20] software. However, PHYLIP has to be downloaded and installed separately; I can not include it into the GENtle package due to licensing issues. GENtle will ask for the location of the PHYLIP binary and treat it transparently after that, i.e., GENtle will call PHYLIP with the user data and display the output without the user noticing the call of an external program.

5.2.5 Sequencer data



```
C: 59;  G: 11;  N: 35
Sample : A
Lane : 1
Start : 2004- 5-11, 20:18:54
Stop : 2004- 5-11, 22:44:26
Gel  :
File :
Primer mobility correction file : DT3100POP6{BDv3}v1.mob
```

Figure 10: The sequencer data module.

The sequence display shows the chromatogram with peaks and the respective color-coded nucleotide. The upper part of the module shows display switches and metadata included with the ABI file.

Today, most DNA sequencing machines function based on the chain termination method [21]. One of the most used formats for storing the chromatograms from these machines is ABI (sometimes AB1). GENtle can natively read this format, and display the chromatogram in its own module (Figure 10). An import filter for the Standard Chromatogram Format (SCF) [22] and ZTR [23] is currently under development.

The sequence can be edited (overwrite only), copied, and stored in a database. I have omitted the storage of complete sequencer data sets, as sqlite databases might clutter from the sheer amount of data (200-400 KByte per data set) over time. This might change once sqlite can store compressed binary data.

Alignments of sequencer data against each other or another sequence can be done manually or via the Sequencing Assistant, which will automatically determine the correct reading direction of each sequencer data set, if applicable. Via the previously described mechanism, the user can go from an alignment mismatch to the respective nucleotide in the chromatogram to manually discern between an altered sequence or a misjudgement by the sequencer software.

5.2.6 Virtual cloning

Virtual cloning is not a module in itself, but describes the process of virtual cutting of DNA strands with restriction endonucleases, and the subsequent ligation of some of the resulting DNA fragments to a new sequence.

A DNA sequence can be subjected to a virtual digestion by restriction endonucleases. An assistant dialog (→5.2.12.2) allows for a preview of the resulting fragments, and their generation as new DNA sequences with blunt or sticky ends. The enzymatic cleavage can also be shown on a virtual agarose gel.

The opposite process, ligation, can be simulated as well (→5.2.12.3). From a list of DNA fragments, an assistant dialog shows the possible resulting DNA sequences, both linear and circular, if applicable. The selected results can be generated as new DNA sequences.

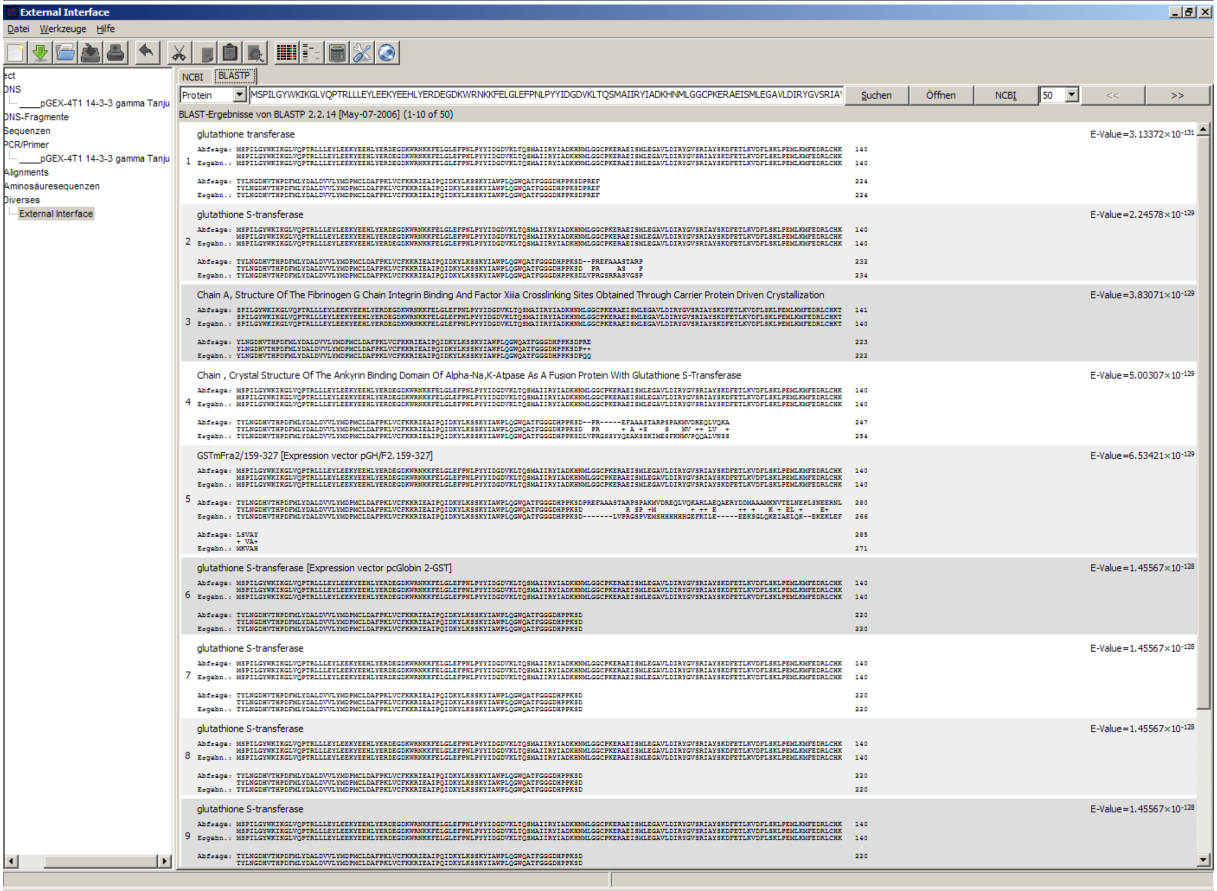
5.2.7 Virtual gels

To visualize or validate the agarose gel resulting from a DNA endonuclease restriction, a virtual gel can be generated that simulates the electrophoresis. The restriction assistant can add a restriction to an existing gel, or create a new one if necessary. Gel lanes can be created as one lane per restriction endonuclease (single enzyme digestion), or with all endonuclease digestion results in a single lane, either fully or partially digested.

An extendable set of DNA ladders is available to run against the virtual samples. The gel can be altered in its virtual concentration. All lane labels are editable, and the agarose concentration can be changed. The gel image can be copied, printed, or saved as an image. The gel display is, on purpose, not photorealistic, to remove any temptation for the user of skipping the actual gel electrophoresis and publishing *in silico* gels instead.

5.2.8 Web interface

The web interface module integrates database searches on the internet with GENtle. Database searches are initiated from within the software, and the results are displayed in a list to chose from.



```
MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIA
BLAST-Ergebnisse von BLASTP 2.2.14 [May-07-2006] (1-10 of 50)

  glutathione transferase                                                                                                                  E-Value=3.13372×10^-131
  Abfrage: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140
           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK
1 Ergebn.: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140

  Abfrage: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDPREF                                                         224
           TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDPREF
  Ergebn.: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDPREF                                                         224

  glutathione S-transferase                                                                                                                E-Value=2.24578×10^-129
  Abfrage: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140
           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK
2 Ergebn.: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140

  Abfrage: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD--PREFAAASTARP                                              232
           TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD  PR    AS   P
  Ergebn.: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDLVPRGSRRASVGSP                                              234

  Chain A, Structure Of The Fibrinogen G Chain Integrin Binding And Factor Xiiia Crosslinking Sites Obtained Through Carrier Protein Driven Crystallization   E-Value=3.83071×10^-129
  Abfrage: SPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHKT   141
           SPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHKT
3 Ergebn.: SPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHKT   140

  Abfrage: YLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDPRE                                                           223
           YLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDP++
  Ergebn.: YLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDPQQ                                                           222

  Chain , Crystal Structure Of The Ankyrin Binding Domain Of Alpha-Na,K-Atpase As A Fusion Protein With Glutathione S-Transferase                           E-Value=5.00307×10^-129
  Abfrage: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140
           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK
4 Ergebn.: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140

  Abfrage: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD--PR-----EFAAASTARPSPAKMVDREQLVQKA                           247
           TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD  PR     + A +S    S   MV ++ LV  +
  Ergebn.: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDLVPRGSSYYQEAKSSKIMESFKNMVPQQALVNSS                           254

  GSTmFra2/159-327 [Expression vector pGH/F2.159-327]                                                                                     E-Value=6.53421×10^-129
  Abfrage: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140
           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK
  Ergebn.: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140
5
  Abfrage: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDPREFAAASTARPSPAKMVDREQLVQKARLAEQAERYDDMAAAMKNVTELNEPLSNEERNL   280
           TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD          R SP +M         + ++ E     ++ +   K + EL +    E+
  Ergebn.: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD-------LVPRGSPVEMSHHHHHHGEFKILE-----EEKSGLQKEIAELQK--EKEKLEF   266

  Abfrage: LSVAY                                                                                                                                      285
           + VA+
  Ergebn.: MKVAH                                                                                                                                      271

  glutathione S-transferase [Expression vector pcGlobin 2-GST]                                                                            E-Value=1.45567×10^-128
  Abfrage: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140
           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK
6 Ergebn.: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140

  Abfrage: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD                                                             220
           TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD
  Ergebn.: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD                                                             220

  glutathione S-transferase                                                                                                                E-Value=1.45567×10^-128
  Abfrage: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140
           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK
7 Ergebn.: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140

  Abfrage: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD                                                             220
           TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD
  Ergebn.: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD                                                             220

  glutathione S-transferase                                                                                                                E-Value=1.45567×10^-128
  Abfrage: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140
           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK
8 Ergebn.: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140

  Abfrage: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD                                                             220
           TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD
  Ergebn.: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD                                                             220

  glutathione S-transferase                                                                                                                E-Value=1.45567×10^-128
  Abfrage: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140
           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK
9 Ergebn.: MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK   140

  Abfrage: TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD                                                             220
           TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD
```

Figure 11: The web interface.

Showing the results of a BLAST search for the amino acid sequence of a Glutathion-S-transferase (GST) protein.

The module can search databases from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) for nucleotide and amino acid sequences by name. Returned sequence entries can be opened via mouse click into fully annotated sequences. Alternatively, the result can be opened on the NCBI web site.

In a similar fashion, PubMed (→4.4.2) can be searched, specifying fields for key words, authors, and year range. The results list shows a preview of the paper title, publication date, journal, and authors. Results can be click-opened in the web browser.

Also, BLAST (→4.4.1) searches for either BLASTN (nucleotides) or BLASTP (protein) run through the web interface (Figure 11), though they are invoked from the DNA or amino acid modules, respectively. BLAST results are shown as a simple alignment of the original and the found match, together with the name of the found match as well as the “E value”, which indicates the similarity of the sequences. A mouse click can open the annotated matching sequence in a new DNA or amino acid sequence module, respectively.

5.2.9 Chromatograms

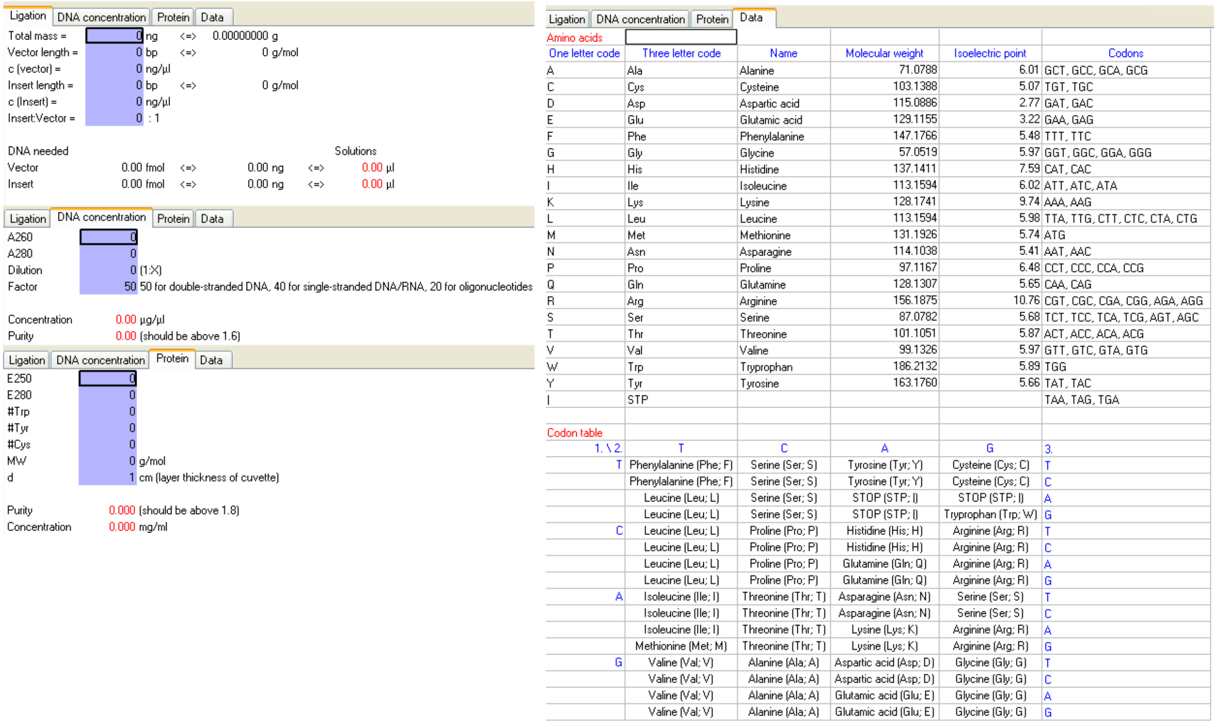
The chromatogram viewer module is a not sequence-based module. It is designed to display data recorded by HPLCs, FPLCs, photometers, fluorimeters, etc. Though this module is still in its early stages of development, it can read and display several data formats, including some photo-/fluorimeter types, some BioRad FPLCs, and raw XY pair as comma-separated values (CSV). The auto-scaling display can be zoomed, and a few visual options can be changed. The current view can be copied to the clipboard, printed, or saved as an image. Planned improvements in this module include automatic peak detection and integration.

5.2.10 Image Viewer

The image viewer module is another visualization module. Seemingly an unlikely function for a molecular biology software, it was born of the necessity to access gel images and phosphoimager scans in an old proprietary format, lacking any description. It also supports a few standard image formats. It can save images in standard formats regardless of their original format, print, or copy them to the clipboard. Planned features include more raw image import filters, as well as a band/plot detection and analysis.

5.2.11 Calculators

Calculators are also not sequence-based. Each calculator is a spreadsheet-based representation of a function often occurring in daily laboratory routine (Figure 12).



Total mass = 0 ng <=> 0.00000000 g
Vector length = 0 bp <=> 0 g/mol
c (vector) = 0 ng/µl
Insert length = 0 bp <=> 0 g/mol
c (Insert) = 0 ng/µl
Insert:Vector = 0 : 1

DNA needed Solutions
Vector 0.00 fmol <=> 0.00 ng <=> 0.00 µl
Insert 0.00 fmol <=> 0.00 ng <=> 0.00 µl

A260 0
A280 0
Dilution 0 (1:X)
Factor 50 50 for double-stranded DNA, 40 for single-stranded DNA/RNA, 20 for oligonucleotides

Concentration 0.00 µg/µl
Purity 0.00 (should be above 1.6)

E250 0
E280 0
#Trp 0
#Tyr 0
#Cys 0
MW 0 g/mol
d 1 cm (layer thickness of cuvette)

Purity 0.000 (should be above 1.8)
Concentration 0.000 mg/ml

Amino acids

One letter code	Three letter code	Name	Molecular weight	Isoelectric point	Codons
A	Ala	Alanine	71.0788	6.01	GCT, GCC, GCA, GCG
C	Cys	Cysteine	103.1388	5.07	TGT, TGC
D	Asp	Aspartic acid	115.0886	2.77	GAT, GAC
E	Glu	Glutamic acid	129.1155	3.22	GAA, GAG
F	Phe	Phenylalanine	147.1766	5.48	TTT, TTC
G	Gly	Glycine	57.0519	5.97	GGT, GGC, GGA, GGG
H	His	Histidine	137.1411	7.59	CAT, CAC
I	Ile	Isoleucine	113.1594	6.02	ATT, ATC, ATA
K	Lys	Lysine	128.1741	9.74	AAA, AAG
L	Leu	Leucine	113.1594	5.98	TTA, TTG, CTT, CTC, CTA, CTG
M	Met	Methionine	131.1926	5.74	ATG
N	Asn	Asparagine	114.1038	5.41	AAT, AAC
P	Pro	Proline	97.1167	6.48	CCT, CCC, CCA, CCG
Q	Gln	Glutamine	128.1307	5.65	CAA, CAG
R	Arg	Arginine	156.1875	10.76	CGT, CGC, CGA, CGG, AGA, AGG
S	Ser	Serine	87.0782	5.68	TCT, TCC, TCA, TCG, AGT, AGC
T	Thr	Threonine	101.1051	5.87	ACT, ACC, ACA, ACG
V	Val	Valine	99.1326	5.97	GTT, GTC, GTA, GTG
W	Trp	Tryprophan	186.2132	5.89	TGG
Y	Tyr	Tyrosine	163.1760	5.66	TAT, TAC
|	STP				TAA, TAG, TGA

Codon table

1. \ 2.	T	C	A	G	3.
T	Phenylalanine (Phe; F)	Serine (Ser; S)	Tyrosine (Tyr; Y)	Cysteine (Cys; C)	T
	Phenylalanine (Phe; F)	Serine (Ser; S)	Tyrosine (Tyr; Y)	Cysteine (Cys; C)	C
	Leucine (Leu; L)	Serine (Ser; S)	STOP (STP; |)	STOP (STP; |)	A
	Leucine (Leu; L)	Serine (Ser; S)	STOP (STP; |)	Tryprophan (Trp; W)	G
C	Leucine (Leu; L)	Proline (Pro; P)	Histidine (His; H)	Arginine (Arg; R)	T
	Leucine (Leu; L)	Proline (Pro; P)	Histidine (His; H)	Arginine (Arg; R)	C
	Leucine (Leu; L)	Proline (Pro; P)	Glutamine (Gln; Q)	Arginine (Arg; R)	A
	Leucine (Leu; L)	Proline (Pro; P)	Glutamine (Gln; Q)	Arginine (Arg; R)	G
A	Isoleucine (Ile; I)	Threonine (Thr; T)	Asparagine (Asn; N)	Serine (Ser; S)	T
	Isoleucine (Ile; I)	Threonine (Thr; T)	Asparagine (Asn; N)	Serine (Ser; S)	C
	Isoleucine (Ile; I)	Threonine (Thr; T)	Lysine (Lys; K)	Arginine (Arg; R)	A
	Methionine (Met; M)	Threonine (Thr; T)	Lysine (Lys; K)	Arginine (Arg; R)	G
G	Valine (Val; V)	Alanine (Ala; A)	Aspartic acid (Asp; D)	Glycine (Gly; G)	T
	Valine (Val; V)	Alanine (Ala; A)	Aspartic acid (Asp; D)	Glycine (Gly; G)	C
	Valine (Val; V)	Alanine (Ala; A)	Glutamic acid (Glu; E)	Glycine (Gly; G)	A
	Valine (Val; V)	Alanine (Ala; A)	Glutamic acid (Glu; E)	Glycine (Gly; G)	G

Figure 12: The calculator module.

The figure is merged from four separate screenshots. On the left, ligation calculator (top), photometric DNA (middle) and protein (bottom) calculator. On the Right, codon and reverse codon table.

At the moment of writing, GENtle includes calculators to

- determine the volumes of vector and insert for ligations, based on their respective length and concentration, as well as the desired end mass and ratio.
- determine the concentration and purity of DNA in solution based on photometric measurements at 260 and 280 nm.
- determine the concentration and purity of protein in solution based on photometric measurements at 250 and 280 nm, as well as the number of tryprophan, tyrosine, and cysteine, respectively. These can be pre-filled automatically by the amino acid sequence module for a specific protein.

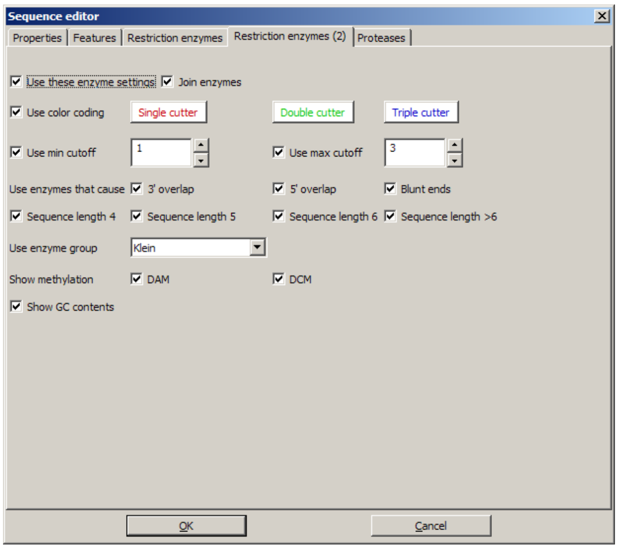
For convenience, I added a codon and reverse codon table to the calculator module, utilizing the spreadsheet layout. New instances of both calculators and data sheets can be added upon request.

5.2.12 Dialogs and assistants

There are several minor and major dialogs and assistant functions present in GENtle, ranging from global program options to simple value requests. Some of them have already been discussed; three more should be mentioned in detail.

5.2.12.1 Sequence editor

The sequence editor allows to edit both DNA and amino acid sequence metadata, including name, description, and active proteases. For DNA sequences, it also allows for the addition of restriction endonucleases, both manually and automatically.



☑ Use these enzyme settings ☑ Join enzymes

☑ Use color coding

☑ Use min cutoff 1 ☑ Use max cutoff 3

Use enzymes that cause ☑ 3' overlap ☑ 5' overlap ☑ Blunt ends

☑ Sequence length 4 ☑ Sequence length 5 ☑ Sequence length 6 ☑ Sequence length >6

Use enzyme group Klein

Show methylation ☑ DAM ☑ DCM

☑ Show GC contents

Figure 13: The sequence editor.

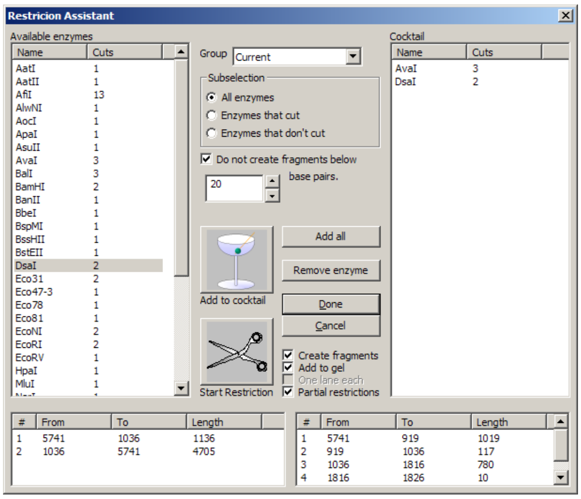
Shown is the automatic restriction endonuclease selector.

The manual restriction endonuclease management lets the user chose a set of restriction endonucleases, whose cuts in the sequence will be displayed if applicable. For convenience, groups of enzymes can be defined and altered, for example, all restriction endonucleases available to the work group. These enzyme groups can be shared via common databases.

Restriction endonuclease cuts can also be displayed for such enzymes that correspond to a set of rules (Figure 13). These rules can be set globally through the program options dialog (not shown), or per sequence, overriding the global settings. The rules include limits for minimal and maximal number of cuts of an enzyme in the sequence, the type of overlap a cut leaves, the length of the recognition sequence, and the choice of enzymes from an enzyme group as mentioned above. Further, it contains display options about the style and color-coding of the enzymes, as well as display options for methylation (DAM and DCM) and the GC contents of the sequence.

5.2.12.2 Restriction assistant

The restriction assistant (Figure 14) allows to simulate the effects of restriction endonucleases on a DNA sequence. Enzymes from a list generated from aforementioned groups and cutting properties can be combined to a restriction "cocktail". The fragments resulting from a digestion with a single enzyme or the cocktail are shown in respective lists.



Restricion Assistant

Available enzymes

Name	Cuts
AatI	1
AatII	1
AfiI	13
AlwNI	1
AocI	1
ApaI	1
AsuII	1
AvaI	3
BalI	3
BamHI	2
BanII	1
BbeI	1
BspMI	1
BssHII	1
BstEII	1
DsaI	2
Eco31	2
Eco47-3	1
Eco78	1
Eco81	1
EcoNI	2
EcoRI	2
EcoRV	1
HpaI	1
MluI	1

Group: Current

Subselection

● All enzymes
○ Enzymes that cut
○ Enzymes that don't cut

☑ Do not create fragments below 20 base pairs.

☑ Create fragments
☑ Add to gel
□ One lane each
☑ Partial restrictions

Cocktail

Name	Cuts
AvaI	3
DsaI	2

#	From	To	Length
1	5741	1036	1136
2	1036	5741	4705

#	From	To	Length
1	5741	919	1019
2	919	1036	117
3	1036	1816	780
4	1816	1826	10

Figure 14: The restriction assistant

Resulting DNA fragments can be generated as new DNA sequence modules, optionally limiting the creation of very short fragments. A virtual gel (→5.2.7) can be generated for the fragments of each enzyme separately or together, optionally for partial digestions as well.

The resulting sequence will be sorted in the main tree under "DNA Fragments" for better separation from source sequences, but do not differ otherwise from other DNA sequences. They will be loaded from the database under the normal DNA heading.

5.2.12.3 Ligation assistant

The ligation assistant can simulate a single- or multipoint ligation. From all non-circular sequences, the assistant runs through the possible ligations and displays the potential products. Sequences and products can be chosen by the user, and the products can be generated as new sequences.

5.3 Platform notes

5.3.1 Windows

GENtle development was initially done on Windows only, attempting to address the large Windows user base before porting it to other platforms. Development of new functions is still mostly done under Windows, mostly for the comfortable development environment offered by DEV-C++ (→4.2.1).

5.3.2 Mac

Despite MYSQL and SQLITE3 being available by default on Mac OSX since version 10.4, I decided to manually compile and statically link libraries of MYSQL, SQLITE2, and SQLITE3 to GENtle, to ensure downward compatibility (early GENtle versions used SQLITE2) and full function spectrum on OSX versions prior to 10.4.

While XCODE (→4.2.2) can generate „universal binaries“ (executable files that run natively on both PowerPC and Intel processors), difficulties with compiling universal binary libraries have prevented me from compiling a GENtle universal binary. Until this point can be resolved, GENtle does run sufficiently under the Rosetta emulation software which is included in the Intel version of OSX 10.4 and above.

5.3.3 Linux

Due to the lack of a standard binary installation method on the different Linux distributions, the Linux binary I offer is often outdated and lacks several steps behind the current development. My recommendation is to download the source code instead, and compile GENtle using the AUTOMAKE/AUTOCONF system. This has proven to be a usable arrangement on various distributions and hardware systems, including 64 Bit processors.

I have initiated contacts with some DEBIAN developers, since many bioinformatics-specific Linux distributions, as well as popular ones like UBUNTU and KNOPPIX, are Debian-based. With their assistance, I hope to be able to include GENtle into the standard package repository of such distributions. Such a package could then be converted into the RPM standard, which is used by distributions such as RedHat and SuSE.

5.4 Concluding discussion

Since the beginning of its development in 2003, GENtle has grown to accommodate the needs of daily lab routine, and continues to prevail in the most thorough test of all, practical use.

My design goal of intuitive use has been achieved, as far as can be told from casual user interviews, as well as from oral and electronic feedback.

The design goal of efficiency through speed has been achieved as well, easily demonstrated by working on practical problems. A favored example of mine is importing the GenBank-formatted file containing the complete genome (2.8 million base pairs) of *Staphylococcus aureus*, which takes about 5 seconds on my laptop, including parsing annotated features and DNA sequence from the GenBank file, translating the DNA of all features with a given reading frame into the corresponding amino acid sequence, checking for restriction endonuclease cuts, as well as layouting and drawing both sequence and map display. Operations on the sequence, like searching, selecting, etc. then occur without noticeably delay. At the same time, the memory consumption (~60MB) is rather moderate when taking the amount of metadata into account.

GENtle does not cover all the possibilities of working with DNA and amino acid sequences, as that was never the intention. Instead, it reduces complexity for the user to cover the vast majority of daily tasks in an easy-to-use environment. It is thus fighting the trend of many (commercial) applications, both inside and outside the field of molecular biology, to define progress as drowning the user in a jungle of functions that are rarely or never used, while at the same time complicating every-day operations.

GENtle is not feature-complete, and it will most likely never reach that state. New functions and modules will be added as need arises, and existing ones will be enhanced. Being open source does ensure both further existence and development.

5.5 Outlook

For the near future, the following additions to GENtle are planned:

- *External modules*, both as locally installed 3rd-party software and as internet services, will be usable from within GENtle through a plug-in system. In contrast to the web-based tools (→4.4.3), modules will work transparently, e.g., the user will not notice that the data is processed outside of GENtle.
- *Chromatogram viewer enhancements* will, as discussed, include more import filters and analysis tools. A spreadsheet-based output for both raw and analysis data will act as an in-between with other software.
- *Sequence versioning and inheritance* will allow for backtracking the origins of a construct through the database. This is currently done through automatically generated comments.
- *File system organizer* will integrate external files (e.g., sequencer data, chromatograms) into the GENtle database-driven sequence management system.

6 References

1 Williams S (2002)
Free as in Freedom
O'Reilly Press

2 Rice P, Longden I and Bleasby A (2000)
EMBOSS: The European Molecular Biology Open Software Suite
Trends in Genetics **16**(6), 276-277

3 Olson GM and Olson JS (2003)
Human-Computer Interaction: Psychological Aspects of the Human Use of Computing
Annu. Rev. Psychol. **54**, 491–516

4 Eric Steven Raymond (2003)
The Art of Unix Programming.
Addison Wesley

5 Needleman S, Wunsch C (1970)
A general method applicable to the search for similarities in the amino acid sequence of two proteins
J Mol Biol. **48**(3), 443-453

6 Hirschberg DS (1975)
A linear space algorithm for computing maximal common subsequences
Comm. A.C.M. **18**(6), 341-343

7 Smith TF and Waterman MS (1981)
Identification of Common Molecular Subsequences
Journal of Molecular Biology **147**, 195-197

8 Thompson JD, Higgins DG, Gibson TJ (1994)
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment
Nucleic Acids Res. **22**(22), 4673-80

9 Lupas, van Dyke and Stock (1991)
Predicting coiled coils from protein sequences
Science **252**, 1162-1164

10 Elbashir S, Lendeckel W, and Tuschl T (2001)
RNA interference is mediated by 21 and 22 nt RNAs
Genes & Dev. **15**, 188-200

11 Chou PY and Fasman GD (1978)
Prediction of the secondary structure of proteins from their amino acid sequence.
Adv Enzymol Relat Areas Mol Biol. **47**, 45-148

12 Kyte J and Doolittle RF (1982)
A simple method for displaying the hydropathic character of a protein.
J Mol Biol. **157**(1), 105-32

13 Hopp TP and Woods KR (1981)
Prediction of protein antigenic determinants from amino acid sequences.
Proc Natl Acad Sci USA. **78**6, 3824-8

14 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990)
Basic local alignment search tool
J. Mol. Biol. **215**, 403-410

15 Knuth D (1976)
Big Omicron and big Omega and big Theta
ACM SIGACT News **8**(2), 18-24

16 Marinus MG and Morris NR (1973)
Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12
J. Bacteriol. **114**, 1143-1150

17 Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005)
The Proteomics Protocols Handbook : Protein Identification and Analysis Tools on the ExPASy Server
Humana Press

18 Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985)
Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis
Science **230**(4732), 1350-1354

19 Allawi HT and SantaLucia J (1997)
Thermodynamics and NMR of internal G-T mismatches in DNA
Biochemistry **36**, 10581-10594

20 Felsenstein J (1989)
PHYLIP - Phylogeny Inference Package (Version 3.2)
Cladistics **5**, 164-166

21 Sanger F, Nicklen S, Coulson AR (1977)
DNA sequencing with chain-terminating inhibitors
Proc Natl Acad Sci USA **74**(12), 5463-5467

22 Dear S and Staden R (1992)
A standard file format for data from DNA sequencing instruments.
DNA Sequence **3**, 107-110

23 Bonfield JK and Staden R (2002)
ZTR: a new format for DNA sequence trace data
Bioinformatics **18**(1), 3-10

7 Appendix

7.1 Acknowledgements

My thanks goes to Prof. Dr. H. W. Klein for the opportunity to develop this software, and for support I received from him.

I thank Prof. Dr. Sabine Waffenschmidt for acting as referee, Prof. Dr. Thomas Wiehe for acting as chairman, and Dr. Kristin Baer for acting as assessor during my disputation.

Also, I thank the members of the work group, past and present, who were brave enough to work with the software, especially in the early stages of development, and for their continued feedback. GENtle wouldn't be as feature-rich and usable as it is without them.

Last not least, my thanks goes to the many people and work groups who sent me bug reports, suggestions for improvements or new functions, or moral support by email. These include, in no particular order: the Department of Biosciences, University of Helsinki, Finland; the Subject of Molecular Genetics, University of Hanover, Germany; the Laboratorie de Biologie Théorique; the Institute of Virology, University of Zurich, Switzerland; the Department of Molecular Biology and Biophysics, Mount Sinai School of Medicine, New York, USA; the Institute of Microbiology, Technical University of Munich, Germany; the Department of Pharmacology, University of Tennessee, USA; the Charite, Berlin, Germany; the Laboratory of Behavioural Neurobiology, Swiss Federal Institute of Technology; the Institute for Molecular Biotechnology, Jena, Germany; the Institute for Experimental Cancer Research, Switzerland; the Institute for Medical Radiation and Cell Research, Würzburg, Germany; the California Institute of Technology, USA; the Department of

Biochemistry, University of Oviedo, Spain; the Department of Cell Biology, GBF, Braunschweig, Germany; the Technical Microbiology AB, Technical University Hamburg, Germany; the Institut for Biomedical Engineering, University Hospital Aachen, Germany; the Department of Anatomy, University of Wisconsin-Madison, USA; the Nanotechnology R&D, Ames Research Center, NASA, USA; the School of Fisheries Sciences, Kitasato University, Japan; the Laboratoire de biologie moléculaire er cellulaire, Université de Neuchâtel, Switzerland; Institute of Biochemistry, University of Düsseldorf, Germany; the Institute of Biology/Experimental Biophysics, Humboldt University Berlin, Germany; the Department of Biochemistry, University of Zurich, Switzerland; the Faculty for Clinical Medicine Mannheim, University of Heidelberg, Germany; the MetProt Research Group, Leiden University, the Netherlands; the Edinger Institute of Neurology, Goethe University Frankfurt/Main, Germany; the Department of Microbiology and Genetics, TU Berlin, Germany; the Dipartimento delle Scienze Biologiche, Universita di Napoli, Italy; the Department of Chemical Engineering, University of California at Berkeley, USA; the Vertebrate Development Laboratory, Cancer Research UK, London; the Rudolf Magnus Institute of Neuroscience, University of Utrecht, the Netherlands; the Department of Molecular and Cell Biology, University of California at Berkeley, USA.

7.2 GENtle manual

The following is a copy from the GENtle online help at http://en.wikibooks.org/wiki/GENtle. As such, figures are small and not numbered nor described in detail.

7.2.1 About

7.2.1.1 Style

- Menus, Buttons etc. are marked ***like this***
- "Double click" always refers to the left mouse button
- "Context menu" always refers to the pop-up menu that appears on clicking with the right mouse button

7.2.1.2 Copyright

- GENtle is ©2004 by Magnus Manske, licensed under GPL
- This manual is ©2004 by its authors, licensed under GFDL

7.2.2 Setup

7.2.2.1 Installation

On Windows, run the file GENtleSetup.exe. It will set up GENtle in a directory of your choice, ans also prepare for a clean deinstallation.

7.2.2.2 Databases

There are two types of databases supported by File-based (sqlite) and MySQL. By default, a file-based database local.db is set up in the installation directory of GENtle. You can (and should, if access limitations prevent you from working with local.db) create new databases and share them with other users of GENtle in your work group (if you want to).
Database management can be found through the menu "Tools/Manage database", under the tab "Databases".

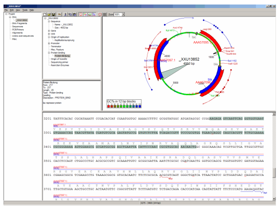
7.2.2.3 File-based databases

You can add an existing or create a new database via the appropriate buttons. Select the database name and location in the following file dialog.

7.2.2.4 MySQL databases

These work similar to the file databases, but require more parameters. Also, a MySQL server has to be installed prior to MySQL database creation. Contact your local system administrator about this. Creating a new MySQL database might require more MySQL privileges than actually using the created database. Thus, after creation, one might want to remove the database entry via the "Remove" button (the database will continue to exist), and add it again with different MySQL privileges.

7.2.3 DNA



The DNA modul.

Within the DNA module, DNA sequences can be viewed, edited and annotated. It is the central module of GENtle. Two major components of the DNA module are the DNA map and the sequence map; see there for details.

7.2.3.1 Open and display DNA

A DNA sequence can be opened in one of the following ways:

- Open from a database
- Import from file
- Manual input
- Create from another DNA module

7.2.3.2 Toolbar

Several functions and display options can be invoked in the tool bar:

- Enter sequence
- Open sequence
- Save sequence
- Undo
- Cut
- Copy
- Paste
- Toggle linear/circular
- Show/hide open reading frames
- Show/hide features
- Show/hide restriction sites
- Expand (=show only) map
- Toggle edit mode
- Zoom

7.2.3.3 Detail tree

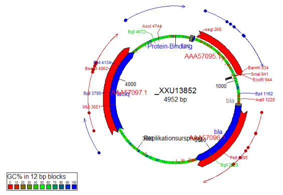
The detail tree, left of the DNA map, shows all parts of the current sequence, including features and restriction enzymes, in a structured fashion. Features and restriction enzymes can be toggled in visibility by a double click, or further manipulated through the context menu.

7.2.3.4 Special menus

View/Show 3'->5'	Show the complementary DNA strand in the sequence map
Edit/Edit ORFs	Change the settings for open reading frame display
Edit/Show possible sequencing primers	Opens the Sequencing primer dialog, which can add possible sequencing primers as features
Edit/Remove sequencing primers	Removes all sequencing primers generated by the above function from the sequence
Edit/Auto-annotate sequences	Finds features from common vectors and other databases in the current sequence
File/Print map	Prints the DNA map
File/Print sequence	Prints the Sequence map
File/Print report	Prints a brief overview. See Printing.

7.2.4 DNA map

The DNA map is shown for DNA sequences (though a variant is also used in protein module for the schematics display). It shows the linear or circular (e.g., plasmid) DNA sequence as a map.



The DNA map.

7.2.4.1 Display

Displayed are

- features (including optional sequencing primers)
- restriction sites
- sticky ends (if any)
- open reading frames (optional)
- sequence name and length (optional)
- methylation (optional)

- GC contents (optional)

7.2.4.2 Mouse actions

Action on	Mouse button	Function
Background	left	Mark sequence
	left (double click)	Open Sequence editor
	middle	Show marked DNA in sequence; show current position in sequence if nothing is marked
Feature	left	Move feature display
	left (double click)	Edit feature (see Sequence editor)
	middle	Mark DNA that matches feature
	middle (shift pressed)	Extend currently marked area to include the DNA of the feature
Restriction site	left	Move site display
	left (double click)	Edit enzyme list (see Sequence editor)
	middle	Open Restriction Assistant
Open reading frame	left	Mark ORF sequence
	left (double click)	Mark and show ORF sequence
All	right	Context menu

7.2.4.3 Context menu

The context menu opens on a click with the right mouse button when somewhere inside the DNA map. The contents of the menu depends on what object in the map you clicked on. Also, depending on the properties of the object, some functions might not be available, for example, amino acids of a feature with no reading frame.

7.2.4.4 Background

Edit sequence	Opens the Sequence editor
Transform sequence	Make sequence inverted and/or complementary
Limit enzymes	Limits enzymes to those that cut no more than ***n*** times
PCR/PCR	Starts the PCR module
PCR/Forward	Starts the PCR module and generates a 5'->3'-primer
PCR/Backward	Starts the PCR module and generates a 5'->3'-primer
PCR/Both	Starts the PCR module and generates both primers
PCR/Mutation	Starts the PCR module and generates overlapping mutagenesis primers
Selection/Cut	Removes the selected part of the sequence and puts it into the clipboard
Selection/Copy	Copys the selected part of the sequence into the clipboard
Selection/Copy to new sequence	Generate a new DNA sequence entry based on the selection
Selection/Show enzymes that cut here	Opens a variant of the Silent Mutagenesis dialog for the selected part of the sequence

Selection/Selection as new feature	Generates a new feature for the selected part of the sequence
Selection/Extract amino acids	Extracts the amino acid sequence of the selected part of the DNA sequence
Selection/BLAST amino acids	Runs a BLAST search for the amino acid sequence of the selected part of the DNA sequence
Selection/BLAST DNA	Runs a BLAST search for the selected part of the DNA sequence
Sequence map/Save as image	Saves the DNA map as an image file
Sequence map/Copy image to clipboard	Copies the DNA map as a bitmap or WMF (see Options) to the clipboard
Sequence map/Print map	Prints the DNA map
Show/hide ORFs	Toggles the open reading frame display
Edit ORFs	Adjusts the open reading frame display

7.2.4.5 Restriction sites

Edit restriction enzyme	Add/remove/manage restriction enzyme via the Sequence editor
Show/hide enzyme	Toggle visibility for the enzyme (this will affect all restriction sites for that enzyme in this sequence)
Remove enzyme	Remove the enzyme from the current selection (this will affect all restriction sites for that enzyme in this sequence). This will ***not*** work for automatically added enzymes (see Options)
Mark restriction site	Marks the recognition sequence of that enzyme at that restriction site
Mark and show restriction site	Marks the recognition sequence of that enzyme at that restriction site and shows it in the sequence
Online enzyme information	Opens the ReBase page for that enzyme
Add to cocktail	This adds the enzyme to the restriction cocktail (see Restriction Assistant)
Add to cocktail	This adds the enzyme to the restriction cocktail (see Restriction Assistant) and starts the restriction

7.2.4.6 Features

Edit feature	Edit the feature properties (see Sequence editor)
Hide feature	Hide the feature from display
Delete feature	Delete the feature
DNA Sequence/Mark feature sequence	Mark the DNA sequence that matches the feature
DNA Sequence/Mark and show feature sequence	Mark the DNA sequence that matches the feature and shows it in the sequence
DNA Sequence/Copy (coding) DNA sequence	Copies the DNA sequence that matches the feature to the clipboard
DNA Sequence/This feature as new sequence	Generates a new DNA sequence based on the feature
DNA Sequence/BLAST DNA	Runs a BLAST search for the DNA of the feature
Amino acid sequence/Copy amino acid sequence	Copies the amino acid sequence of the feature to

	the clipboard
Amino acid sequence/As new entry	Generates a new protein entry based on the amino acid sequence of the feature
Amino acid sequence/Blast amino acids	Runs a BLAST search for the amino acid sequence of the feature

7.2.4.7 Open reading frames

As new feature	Generate a new feature from the ORF, with the appropriate reading frame and direction
DNA sequence/Copy DNA sequence	Copies the DNA sequence of the ORF to the clipboard
DNA sequence/As new DNA	Generates a new DNA sequence entry based on the DNA sequence of the ORF
DNA sequence/BLAST DNA	Runs a BLAST search for the DNA sequence of the ORF
Amino acid sequence/Copy amino acid sequence	Copies the amino acid sequence of the ORF to the clipboard
Amino acid sequence/As new AA	Generates a new protein entry based on the amino acid sequence of the ORF
Amino acid sequence/BLAST amino acids	Runs a BLAST search for the amino acid sequence of the ORF

7.2.5 Sequence map

The sequence map is used by most GENtle modules. It shows sequences of DNA or amino acids, as well as primers, features, restriction sites and more. The basic behaviour, however, is always similar.

The sequence map.

7.2.5.1 Clicks

A double click usually opens the Sequence editor for the sequence.

7.2.5.2 Context menu

The available functions in the context menu vary with the module the sequence map is used in, its state, and selection.

Edit sequence	Turn on edit mode
Transform sequence	Invert and/or complement the sequence (DNA module only)
Limit enzymes	Limit enzymes so that only enzymes below a certain number of cuts in the

	sequence is shown (DNA module only)
PCR	Compare DNA map
Selection	Compare DNA map
Copy as image	Copies the sequence map as a bitmap to the clipboard (***Caveat :*** Such a bitmap can take up a huge amount of memory, depending on the length of the sequence)
Save as image	Saves the sequence map in one of several image formats
Print sequence	Prints the sequence

7.2.5.3 Keys

The whole sequence can be marked by Ctrl-A. The Find dialog can be invoked by Ctrl-F. Both functions can also be called upon through a menu.

In the DNA and PCR modules, the amino acid reading frame can be toggled by keys like this:

- Ctrl-1 = reading frame 1
- Ctrl-2 = reading frame 2
- Ctrl-3 = reading frame 3
- Ctrl-4 = reading frame 1, complementary strand
- Ctrl-5 = reading frame 2, complementary strand
- Ctrl-6 = reading frame 3, complementary strand
- Ctrl-7 = all reading frames, one-letter code
- Ctrl-8 = known reading frames only (from the features)
- Ctrl-0 = hide amino acids
- Ctrl-W = three-letter code (not when displaying all reading frames)
- Ctrl-Q = one-letter code

7.2.5.4 Edit mode

Display and edit mode can be toggled by F2 or the toolbar. During editing, the sequence display is maximized, and the DNA map is hidden, improving ease of edit. Depending on the current module, only some keys are allowed (in the DNA module, "A", "C", "G", and "T") by default; any other key will trigger a request to allow all keys for that sequence, for that session. The cursor can be moved similar to that in a text editor. Insert and overwrite mode can be toggled, except for some modules like

PCR or Sequencing, where overwrite mode is mandatory. In these modules, backspace and delete are disabled as well.

When editing a primer in PCR mode, the "." key copies the base at the current position from the 3'→5' or 5'→3' sequence, respectively.

7.2.5.5 Horizontal mode

In some modules, the sequence display can be toggled to horizontal. This can enhance visibility. Printing, however, is always done in standard ("vertical") mode.

7.2.6 Protein

In this module, amino acid sequences (peptides/proteins) can be viewed, edited and annotated. It uses a sequence map as main display, and a multi-purpose overview display at the top.

The protein module.

7.2.6.1 Toolbar

Several functions and display options can be invoked in the tool bar:

- Enter sequence
- Open sequence
- Save sequence
- Print sequence
- Undo
- Cut
- Copy
- Paste
- Plot (shows a plot within the sequence map)
- Horizontal mode

7.2.6.2 Function display

The smaller display on the top can show several types of information:

Data	Shows some basic data that has been calculated from the sequence
Description	Shows the sequence description

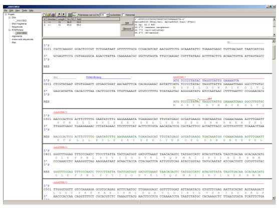
Scheme	Shows a DNA map-like layout of the whole protein
AA weight	Shows a plot of the molecular weight of the individual amino acids
AA isoelectric point	Shows a plot of the isoelectric point of the individual amino acids
Hydrophobicity	Shows a plot of the local hydrophobicity of the amino acids nearby
Chou-Fasman plot	Shows a detailed Chou-Fasman-plot

7.2.6.3 Special menus

Edit/Photometer analysis	Invokes the respective calculator
Edit/'Backtranslate' DNA	Attempts to generate the DNA sequence which codes for this amino acid sequence, using the full range of IUPAC base letters
Edit/Proteolysis assistant	Invokes the Proteolysis assistant

7.2.7 PCR and Primer Design

This module allows for designing primers and running virtual PCRs. It can be started from a DNA module via context menu of the DNA or sequence map, or through ***Tools/PCR***. If a sequence is selected in the DNA module, one or more primers can be generated automatically upon startup of the PCR module. These will just be suggestions, and are in no way optimized by default.



PCR and primer design.

7.2.7.1 Toolbar

- Enter new primer
- Open primer/sequence
- Print PCR
- Add a primer (you will have to open or enter the primer first)
- Export a primer (generate its sequence)
- Edit mode
- Show/hide features
- Polymerase running length
- Horizontal mode

The polymerase running length is the number of nucleotides the polymerase is allowed to run during the PCR in the elongation step. This is usually measured in minutes, but each polymerase runs at a different speed, which is why this information is given here in nucleotides. The value is initially computed automatically, but can be changed manually.

7.2.7.2 Primer list

The primer list (the upper left) shows all primers used in this PCR, as well as certain key properties of these. Selecting one of these primers will show more detailed information in the box on the right (see "Edit primer dialog" for details). Double-clicking one of the primers will mark and show that primer in the sequence. A selected primer can be removed through the ***Remove*** button, or edited via the ***Edit*** button. A selected primer can also be exported via the Export button in the toolbar; a new sequence will be generated for that primer.

- ***Caveat :*** The generated sequence is ***not*** stored anywhere automatically, it needs to be saved manually!
- ***Caveat :*** To add a primer, use the Add button in the toolbar, or the ***Selection as new primer*** context menu. Merely editing the sequence (see below) is for editing existing primers only, it will ***not*** create new ones!

7.2.7.3 Sequence

The sequence consists of the following lines:

- Features of the template DNA (can be turned off in the toolbar)
- 5' primer
- Template DNA (5'→3')
- Amino acid sequence of the template
- Template DNA (3'→5')
- 3' primer
- Restriction sites of the resulting DNA
- Resulting DNA (shown in green)
- Amino acid sequence of the resulting DNA

Some special functions and properties of the PCR sequence display:

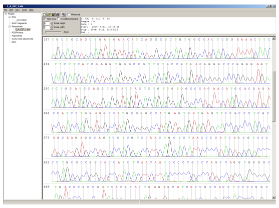
- The amino acid reading frame can be set as described here. This will affect both amino acid sequences shown (template and result).

- Only the two primer sequences can be edited; overwrite mode is mandatory, and deleting is disabled.
- To delete a nucleotide, overwrite it with Space.
- The "." key will copy the matching template nucleotide to that position in the primer sequence.
- Matching primer nucleotides (that is, matching with the template) are shown in blue, mismatches in red.
- If (when ***not*** in edit mode) an empty span of the primer sequence is selected, it can be turned into a new primer via the context menu (***Selection as new primer***).
- The sequence of a restriction site can be inserted left or right of a selection (in edit mode, right or left of the cursor) via the context menu. A selection dialog for the desired enzyme will appear.
- A silent mutation can be introduced via the context menu.

Finally, the resulting DNA or amino acid sequence (the green sequence, which will be the one generated by the PCR) can be copied to the clipboard or generated as a new sequence (containing all features, restriction enzymes etc.) via the context menu.

7.2.8 Sequencing

The sequencing module allows to view the data recorded by a sequence analyser. The data is loaded by importing the appropriate .abi/.ab1 file.



The sequencing module

7.2.8.1 Display

The data is displayed in the main sequence window. The text window on the upper right shows data stored in the file. On the left side, the following display options for the sequence are available:

Help lines	Gray vertical lines down from each sequence letter to the baseline. These can help to identify which letter belongs to which peak

Invert&complement	Shows the sequencing complement/inverted. Useful for alignments
Scale height	Sets the height of the graphic display [unit in text lines]
Scale width	Sets the graphical points per data value. Default is 2; 1 would mean one pixel width per data point
Zoom	Sets the zoom factor for the data; useful to see small peaks

7.2.8.2 Toolbar

- Enter new sequence
- Open sequence
- Save sequence (see caveats)
- Copy sequence to clipboard
- Horizontal mode

7.2.8.3 Caveats

- Editing works in overwrite-mode only
- Saving will only store the ***sequence*** in the database, not the sequencer data (the peaks), due to memory concerns.

7.2.9 Alignments

The alignment module displays alignments of DNA and amino acid sequences. It can be invoked through ***Tools/Alignment*** or Ctrl-G.

Alignment.

7.2.9.1 Settings dialog

The settings dialog will be invoked upon starting the module, or through the "settings" button in the toolbar. The sequences to align, their order, and the alignment algorithm and its parameters can be chosen here. The following algorithms are available:

Clustal-W	This (default) algorithm generates alignments of high quality, but is rather slow for simple alignments, and sometimes stumbles over local alignments.
Smith-Waterman	An internal, fast, but simple algorithm for local alignments, that is, aligning

one or multiple short sequences againast a long one. The long sequence has to be the first. It works great for checking alignments against the expected sequence.

Needlemann-Wunsch An internal, fast, but simple algorithm for global alignments, that is, aligning sequences of roughly the same length (e.g., different alleles of a gene). As with Smith-Waterman, all alignments are made against the first sequence.

Caveat : Clicking ***OK*** in this dialog will recalculate the alignment; the previous alignment and all manual changes made to it will be lost.

7.2.9.2 Toolbar

Several functions and display options can be invoked in the tool bar:

- Enter sequence
- Open sequence
- Save sequence
- Print sequence
- Settings
- Horizontal mode
- Middle mouse button function

7.2.9.3 View menu

Some display options can be combined with each other:

- Bold (shows characters in bold)
- Mono (black-and-white mode)
- Conserved (shows characters that match the one in the first line as dots)
- Identity (toggles the "identity" line)

Some of them exclude one another:

- Normal (shows colored text on white background)
- Inverted (shows white text on colored background)

Some other display options are planned, but not implemented as of now.

7.2.9.4 Sequence display

The sequence map can be altered through the context menu. These changes will only alter the display, ***not*** recalculate the alignment.

- Lines can be moved up or down
- Features for each line can be shown or hidden. By default, features for the first line are shown, features of the other lines are hidden.
- Gaps can be inserted or deleted, in this line, or all except this line. One of these four possible functions is additionally assigned to the middle mouse button; this setting can be changed in the toolbar.
- A double click on a character (***not*** on a gap) opens the "source" window for that sequence (if available), marks and shows the position that was clicked in the alignment. This can be helpful for checking a sequencing.
- Sequences can be marked across multiple lines, then formatted via the ***Appearance*** context menu.

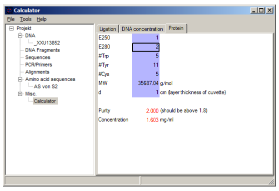
Sequences can ***not*** be edited within the alignment module. For that, you will have to edit the original sequence, then re-run the alignment.

7.2.9.5 Notes

- Legend for the ClustalW consensus line:
- *** = identical or conserved residues in all sequences in the alignment
- **:** = indicates conserved substitutions
- **.** = indicates semi-conserved substitutions

7.2.10 Calculators

The calculator module can be invoked via ***Tools/Calculator***. It contains several specialized spreadsheet-based calculators for typical tasks in molecular biology. The editable fields are shown in blue, the (major) results of the calculation are shown in red.



E250	1	
E280	2	
#Trp	5	
#Tyr	11	
#Cys	5	
MW	35687.04	g/mol
d	1	cm (layer thickness of cuvette)
Purity	2.000	(should be above 1.8)
Concentration	1.603	mg/ml

Protein calculator

7.2.10.1 Ligation calculator

This calculator gives the amount (in µl) of vector and insert for a ligation, based on the length and concentration of each respectively, their desired ratio and total mass of DNA. A typical ratio of insert:vector is 4:1 or 5:1.

7.2.10.2 DNA concentration calculator

This calculator gives the amount and purity of DNA based on photometric absorption at 260 and 280 nm, respectively, as well as the dilution (in case one measures a 1:100 dilution of the original DNA sample) and a correction factor for different types of nucleic acids.

7.2.10.3 Protein calculator

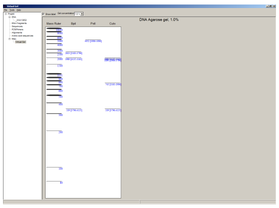
This calculator gives the amount and purity of peptides/proteins based on photometric absorption at 250 and 280 nm, respectively, as well as the molecular weight of the peptide, the layer thickness of the cuvette used, and the number of tryptophans, tyrosines and cysteins in the peptide.

7.2.10.4 Data

This shows a codon table and a reverse codon table, both for standard code. This page can not be edited.

7.2.11 Virtual Gel

A ***virtual*** agarose (DNA) ***gel*** can be generated or expanded via the Restriction assistant. A new virtual gel will be created if none exists; otherwise, the existing one will be expanded.



Virtual gel.

Within the gel viewer, gel concentration can be varied. Also, labelling can be turned on/off. Gels can be printed, or saved/copied as an image.

The name of a lane can be changed by double-clicking on the lane.

You can chose one of several DNA markers. If your favourite marker is not there, you can file a feature request or add your own directly.

7.2.12 Image Viewer

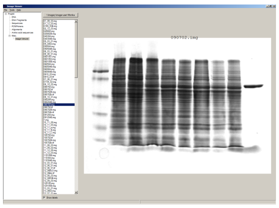


Image viewer.

The Image Viewer module can be invoked via ***Tools/Image viewer***. It can display images, such as gel photos, print them, or save them in another image format.

The viewer can read and write common formats, such as JPG, TIF, BMP, GIF, etc. In addition, it can read the IMG format used by the BioRad Molecular Analyst software.

The directory can be selected via the upper left button. The files in that directory are shown below. A single click on a file displays the image.

The context menu of the image contains entries to save or print the image, or copy it to the clipboard. For saving, PNG, TIF, BMP, and JPG are available formats, with PNG being the default, as it has the best lossless compression.

Labels of IMG images are shown on screen, print, and saved images by default. This can be changed through the "Show labels" checkbox beneath the file list.

An image can be inverted (black <=> white) through the "Invert" checkbox.

7.2.13 Web interface

The GENtle web interface lets you access DNA and amino acid sequences from NCBI, as well as publications listed at PubMed. The interface also covers BLAST searches.

7.2.13.1 NCBI

Choosing ***Nucleotide*** or ***Protein***, entering a sequence name/keywords, and hitting ***Search***/ENTER will show the NCBI search results for that query. More results (if any) can be browsed with >>.

Double-clicking an entry will download and open the (annotated) sequence.

7.2.13.2 PubMed

The ***PubMed*** option gives new entry fields for author(s) (written "Last_name Initials", separated by ","), and date limitations (years), as well as a result sort option.

Double-clicking an entry will open a web browser window with the respective PubMed abstract page.

7.2.13.3 BLAST

Running a BLAST search for a DNA or amino acid sequence will open a new tab in the web interface, showing a countdown for the time the BLAST results are expected to arrive. Once loaded, the results are displayed as simple alignments.

Double-clicking an entry will open the found sequence.

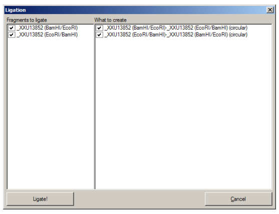
7.2.14 Graph

The ***graph*** module is still in heavy development. It will allow for display of "plotted" data, such as HPLC/FPLC data sets.

7.2.15 Ligation

The ligation dialog is a means for virtually ligating two (or more) DNA fragments. It can be invoked via ***Tools/Ligation*** or Ctrl-L.

The left list shows all potential DNA sequences to be ligated. Some of these are automatically selected, but selection can be manually changed. The right list shows the possible products of a ligation of the selected sequences. Some circular products will be shown in two forms (A-B and B-A), which only differ visually.



Fragments to ligate

☑ _XXU13852 (BamHI/EcoRI)
☑ _XXU13852 (EcoRI/BamHI)

What to create

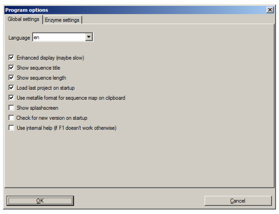
☑ _XXU13852 (BamHI/EcoRI)-_XXU13852 (EcoRI/BamHI) (circular)
☑ _XXU13852 (EcoRI/BamHI)-_XXU13852 (BamHI/EcoRI) (circular)

Ligation dialog.

The selected products will be generated as new sequences on clicking the ***Ligate*** button.

7.2.16 Options

Global program options can be altered via ***Tools/Options***.



Language en

☑ Enhanced display (maybe slow)
☑ Show sequence title
☑ Show sequence length
☑ Load last project on startup
☑ Use metafile format for sequence map on clipboard
☐ Show splashscreen
☐ Check for new version on startup
☐ Use internal help (if F1 doesn't work otherwise)

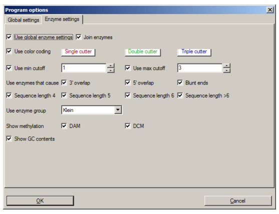
Global settings.

7.2.16.1 Global settings

Option	Description
Language	Currently English, German, and Chinese are available
Enhanced display	Can be turned off on machines with very show graphics
Show sequence title	Displays the sequence title in the DNA map
Show sequence length	Displays the sequence length in the DNA map
Load last project on startup	Automatically loads the last used project when starting GENtle
Use metafile format	Generates a WMF when copying the DNA map instead of a bitmap
Show splash screen	Shows the GENtle splash screen when starting
Check for new version on startup	Checks (and downloads) a new GENtle version via internet on startup
Use internal help	Help should open in a browser window by default. If that doesn't work, check this option

7.2.16.2 Enzyme settings

Here the goal enzyme options can be selected. These can be overridden for an individual sequence in the sequence editor, where there is a tab identical to this one.



☑ Use global enzyme settings ☑ Join enzymes
☑ Use color coding
☑ Use min cutoff 1 ☑ Use max cutoff 3
Use enzymes that cause ☑ 3' overlap ☑ 5' overlap ☑ Blunt ends
☑ Sequence length 4 ☑ Sequence length 5 ☑ Sequence length 6 ☑ Sequence length >6
Use enzyme group Klein
Show methylation ☑ DAM ☑ DCM
☑ Show GC contents

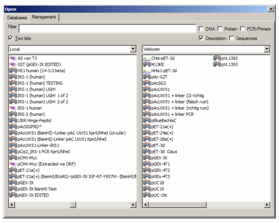
Enzyme settings.

Option	Description
Use global enzyme settings	Turn most of the other options on this tab on or off globally
Join enzymes	In a DNA map, cuts of isoenzymes can be grouped together instead of displayed individually
Use color coding	Restriction enzymes can be shown in a color matching their number of cuts in a given sequence. The three buttons to the right of this option each hold a color choice dialog for single, double, and triple cutters.
Use min/max cutoff	Shows only enzymes that cut a minimum/maximum times
Sequence length	Shows only enzymes with recognition sequences of the selected lengths
Use enzyme group	Uses only enzymes from the selected enzyme group
Show methylation	Shows DAM and/or DCM methylation in map and sequence, in red
Show GC contents	shows the GC contents in the map

7.2.17 Databases

The GENtle database management dialog is where sequences are stored and retrieved. DNA and amino acid sequences, primers, alignments, and projects all go to databases, which can be local (for one computer only) or shared (used by the whole work group).

7.2.17.1 Management



AS von T3
GST (pGEX-3X EDITED)
HS1 human (14-3-3 beta)
IRS-1 (human)
IRS-1 (human) TESTING
IRS-1 (human) UGH!
IRS-1 (human) UGH! 1 of 2
IRS-1 (human) UGH! 2 of 2
IRS-1 human
IRS-2 (human)
LIRK Hinge-Peptid
pAcSIGFRD*
pAcUW51 (BamHI)-!Linker pAC UW51 KpnI/NheI (circulär)
pAcUW51 (BamHI)-Linker pAC UW51 KpnI/NheI
pAcUW51-Linker-IRS1
pCis2_IRS-1 PCR KpnI/NheI
pCMV-Myc
pCMV-Myc (Extracted via ORF)
pET-21a(+)
pET-21a(+) (BamHI/EcoRI)-!pGEX-3X IGF-NT-Y957W- (BamHI/E
pGEX-3X
pGEX-3X BamHI-Test
pGEX-3X EDITED

CHis-pET-3d
M13KE
NHis1-pET-3d
pAc-G2T
pAcSG2
pAcUW51
pAcUW51 + linker (3) richtig
pAcUW51 + linker (falsch rum)
pAcUW51 + linker (richtig rum)
pAcUW51 + linker PCR
pBlueBacHisC
pET-21a(+)
pET-24a(+)
pET-28a(+)
pET-3d
pET-3d Claus
pGEX-3X
pGEX-4T1
pGEX-4T2
pGEx-4T3
pUC18
pUC19
pUC-19c
pVL1392
pVL1393

Databases.

The "Management" tab can be reached through the ***File*** menu, the ***Tools/Manage database*** menu, the Ctrl-O and Ctrl-S keys ("open" and "save", respectively), or the appropriate buttons in the toolbar. The tab consists of two or three parts:

7.2.17.2 Filter

The filter section allows to filter the database entries so the list(s) below show only the matching entries.

The filter text box limits the shown sequences to those whose name (or description or sequence, depending on the checkboxes) contain that text. Multiple search words are separated by a space ("") and work as a logical AND. Thus, entering "pgex igf" in the filter text box shows only those sequences whose name (or description) contain both the word "pgex" and "igf". The search in not case-sensitive, so searching for "igf" or "IGF" will make no difference.

The checkboxes on the right limit the display to any combination of DNA, protein (amino acid sequences), and primers. If non of these is selected, all types of entries are shown, including alignments. As already described, search for text can be extended beyond the sequence name to description and the sequence itself through two other checkboxes, where description search is enabled as default.

7.2.17.3 Lists

One or two lists are shown, depending on the appropriate checkbox above the left list. The database(s) to search/display can be selected via the drop-down box(es). One list with full width is good for an overview of a single database, whereas two lists are needed for moving and copying entries between databases; also, a search will be run on both databases simultaneously.

Entries will be sorted alphabetically. Every entry has a small icon associated with its type. There are icons for DNA, amino acid sequences, primers, and alignments. There is also a project icon, but these will only be shown when opening/saving a project.

A single entry can be selected by clicking with the left mouse button. When opening a file, a double click or pressing RETURN on a selected entry will open it. Multiple entries can be selected by dragging a rectangle with the mouse, or by holding down the SHIFT and/or CTRL keys. A multiple selection can be opened via RETURN.

Grabbing selected entries with the left mouse button and dragging them into the other list will ***move*** these entries to that database. To ***copy*** these entries, hold down the CTRL key when releasing the left mouse button over the target list.

Selected entries can be opened, renamed, and deleted via their context menu.

7.2.17.4 Save

If you save an entry to a database, there will be an additional line below the lists. It consists of a drop-down box with the database to save the entry to, and a text box for the name. The name of the database is remembered if you originally opened that entry from a database, otherwise the standard database is the default.

Saving an entry to a database where an entry with that name already exists will lead to the following:

- If the sequence of the entry in the database is exactly the same as the sequence of the entry you're trying to save, a message box will ask you if you really want to overwrite that entry.
- If the sequence of the entry in the database differs from the sequence of the entry you're trying to save, a message box will tell you that this action was prevented. This will avoid accidental overwriting of an entry with a different sequence. If you are very certain you want to replace that entry, you will have to delete the entry in the database manually via the context menu, as described above.

7.2.17.5 Databases

Currently, GENtle supports sqlite and MySQL databases, both of them freely available. Each has different advantages and disadvantages, though both are integrated seamlessly into GENtle. Once set up, all functions are available on all databases, no matter the type.

The "Databases" tab keeps a list of all the databases that can be accessed. New databases can be created, and existing can be added to or removed from that list. The exception is the local database, which is essential for the functioning of GENtle and therefore can not be removed. Removal of a database will ***not*** delete the database itself, only the entry in the list.

One of the databases in the list is the default database. The default database can be set by selecting its entry in the list, then clicking the ***As Default*** button. The default database can carry shared enzyme groups.

7.2.17.6 Sqlite

Sqlite is already integrated in GENtle, so no separate installation or setup of any kind is required. A sqlite database consists of a single file with the ending ".db". For each GENtle installation, a database ("local.db") is automatically created. New sqlite databases can be created, or existing ones added to GENtle, on the "Databases" tab in the dialog. To take such a database with you (e.g., for use at home or on a laptop), just copy the ".db" file. While sqlite databases are easy to set up and maintain, sharing them across a network tends to be slow, depending on the size of the database.

7.2.17.7 MySQL

MySQL is a professional client/server database system that will reliably store and serve millions of entries. It is ideal for shared databases, as even a huge number of stored sequences will not slow it down significantly, even across a network. However, there are some steps required to use MySQL databases with

- A "server" computer on your network, that is, a computer that is running most of the time, and preferably is not used for direct work. If the server is not running, or disconnected from the network, no one will be able to access the MySQL database and the sequences stored in it!
- The MySQL server software (4.1 works fine, other versions will likely do as

well), which available for free here.

- Someone to configure the MySQL server (not as complicated as it sounds)

Once the MySQL setup is complete, MySQL databases can be created (by one) and added to all the GENtle clients that should have access.

7.2.18 Import

The import dialog is a standard "file open" dialog. It can be invoked via ***Files/Import*** or Ctrl-I.

Multiple files can be chosen to be imported in a row. GENtle will automatically try to determine the file type, but also a file type can be chosen manually.

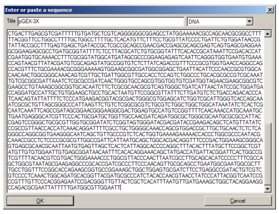
Supported formats include:

- GenBank
- GenBank XML
- FASTA
- ABI/AB1 (popular sequencer output format)
- PDB (a 3D format, import as annotated sequence)
- Clone (old DOS program, proprietary format)
- Numerous other sequence formats that will be imported as "sequence only", without annotations, features etc.
- Comma-separated values (CSV) from various machines (HPLC/FPLC/photometer) for Graphs

7.2.19 Enter sequence

This dialog to enter a sequence manually can be invoked via ***File/Enter sequence*** or Ctrl-N.

Beside the sequence, to be typed or pasted into the large text box, one can enter a title (name) for that sequence, and choose a type.



Title: pGEX-3X

DNA

```
CTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTT
TTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCG
TATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAA
GCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATAAATTCCGACACCAT
CGAATGGTGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCAGGGTGGTGAATGTGAAA
CCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAACCAGGCCAG
CCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCA
CAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAAT
TGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTC
GAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGA
CCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCA
TCAACAGTATTATTTTCTCCCATGAAGACGGTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAA
ATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCG
CAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAAATGC
TGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTAC
CGAGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTATATC
CCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCA
GGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACG
CAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCA
GTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGT
ATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGGATTCACTGGCCG
TCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCA
GCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGCTT
TGCCTGGTTTCCGGCACCAGAAGCGGTGCCGGAAAGCTGGCTGGAGTGCGATCTTCCTGAGGCCGATACTGTCGTC
GTCCCCTCAAACTGGCAGATGCACGGTTACGATGCGCCCATCTACACCAACGTAACCTATCCCATTACGGTCAATCCG
CCGTTTGTTCCCACGGAGAATCCGACGGGTTGTTACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGG
CCAGACGCGAATTATTTTTGATGGCGTTGGAATT
```

Enter a sequence manually.

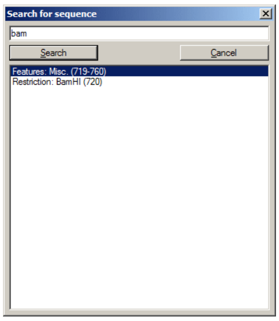
Available types include:

- DNA
- Amino acid sequence
- GenBank
- (GenBank) XML
- Primer

When choosing DNA, amino acids, or primer, all non-sequence characters, like blanks and numbers, are automatically removed.

- ***Note :*** A primer ***has*** to be given the type "Primer", otherwise it will be added as DNA.

7.2.20 Find dialog



bam

Features: Misc. (719-760)
Restriction: BamHI (720)

Find dialog.

The ***Find dialog*** in DNA and amino acid sequence can be invoked via Ctrl-F or ***Edit/Search***. It displays can find a string in

- the current sequence
- a feature name
- a feature description

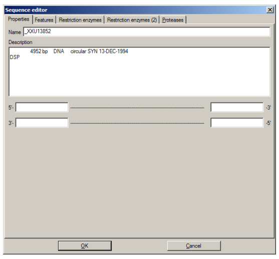
In DNA sequence display, it also look in

- the reverse sequence
- the translated amino acid sequence(s)
- restriction enzyme names

The search is commenced automatically after changing the search string, if it is three or more characters long. For shorter search queries, the ***Search*** button has to be clicked. Single-clicking on a search result will select and display the result in the sequence. A double click will exit the dialog, and open the sequence editor for features, or the enzyme management dialog for restriction enzymes.

7.2.21 Sequence editor

The sequence editor holds the key to several properties of a sequence. It consists of several tabs, depending on the type of sequence, which can be DNA or amino acid.



Name: _XXU13852

Description:
4952 bp DNA circular SYN 13-DEC-1994
DSP

Sequence editor, properties tab.

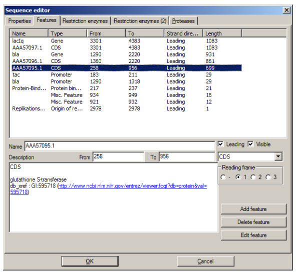
7.2.21.1 Properties

Here, the title and description of the sequence can be altered. As for feature descriptions, the sequence description will make http references clickable. For DNA sequences, sticky ends can be entered.

7.2.21.2 Features

This tab shows a list of all features of the sequence. Features can be added, edited, and deleted. Most of the settings should be self-explanatory.

- The setting ***reading frame*** is only available when the type is set to "CDS" ("coding sequence").
- A ***leading*** sequence is read 5'→3'; leading unchecked, 3'→5'
- ***Edit feature*** will invoke an additional "Edit feature" dialog



Name	Type	From	To	Strand dire...	Length
lacIq	Gene	3301	4383	Leading	1083
AAA57097.1	CDS	3301	4383	Leading	1083
bla	Gene	1290	2220	Leading	931
AAA57096.1	CDS	1360	2220	Leading	861
AAA57095.1	CDS	258	956	Leading	699
tac	Promoter	183	211	Leading	29
bla	Promoter	1290	1318	Leading	29
Protein-Bind...	Protein bin...	217	237	Leading	21
	Misc. Feature	934	949	Leading	16
	Misc. Feature	921	932	Leading	12
Replikations...	Origin of re...	2978	2978	Leading	1

Name: AAA57095.1 ☑ Leading ☑ Visible

Description From: 258 To: 956 CDS

Reading frame: ○ - ● 1 ○ 2 ○ 3

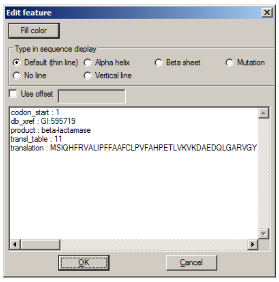
CDS

glutathione S-transferase
db_xref : GI:595718 (http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&val=595718)

The sequence editor, features tab.

7.2.21.3 Edit feature

- ***Fill color*** is the color of the feature; it will invoke a color choice dialog
- ***Type in sequence display*** determines how that feature is drawn in the DNA map
- ***Use offset*** sets the numbering for the first amino acid of the feature; useful if the feature marks a part of a protein



Type in sequence display: ● Default (thin line) ○ Alpha helix ○ Beta sheet ○ Mutation ○ No line ○ Vertical line

□ Use offset

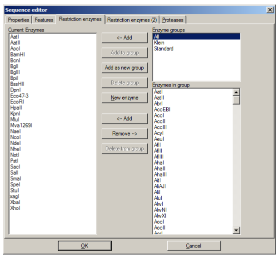
codon_start : 1
db_xref : GI:595719
product : beta-lactamase
transl_table : 11
translation : MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGY

Edit feature dialog.

The list box below contains original data from GenBank format import.

7.2.21.4 Restriction enzymes

When editing a DNA sequence, two tabs with settings for restriction enzymes are available. The first one is identical to the enzyme management dialog. The second one is identical to the global enzyme settings tab, but contains the settings for this sequence alone. By default, its options are disabled, and the global options are used. By activation the options here, global settings are overridden.



Current Enzymes: AatI, AatII, AocI, BamHI, BcnI, BglI, BglII, BpiI, BssHII, DpnI, Eco47-3, EcoRI, HpaII, KpnI, MluI, Mva1269I, NaeI, NcoI, NdeI, NheI, NotI, PstI, SacI, SalI, SmaI, SpeI, StuI, xagI, XbaI, XhoI

Enzyme groups: All, Klein, Standard

Enzymes in group: AatI, AatII, AbrI, AccEBI, AccI, AccII, AccIII, AcyI, AeuI, AflI, AflII, AflIII, AhaI, AhaII, AhaIII, AitI, AliAJI, AliI, AluI, AlwI, AlwNI, AlwXI, AocI, AocII

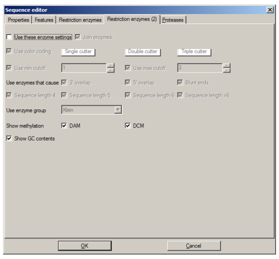
Sequence editor, enzyme settings tab.

7.2.21.5 Proteases

This tab holds a list of available proteases. Potential cleavage sites for selected (checked) proteases are shown in the sequence (***not*** in the DNA map).

New proteases can be added similar to the following examples:

- Example : "Thermolysin"
  - Sequence for this protease : "!DE|AFILMV"
  - Explanation: "Not D or E", "cut", "then A, F, I, L, M, or V"
- Example: Prolin-endopeptidase
  - Sequence for this protease : "HKR,P|!P"
  - Explanation : "H, K, or R", "then P", "cut", "then not P"



☐ Use these enzyme settings ☑ Join enzymes

☑ Use color coding

☑ Use min cutoff 1 ☑ Use max cutoff 3

Use enzymes that cause ☑ 3' overlap ☑ 5' overlap ☑ Blunt ends

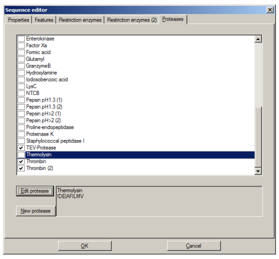
☑ Sequence length 4 ☑ Sequence length 5 ☑ Sequence length 6 ☑ Sequence length >6

Use enzyme group Klein

Show methylation ☑ DAM ☑ DCM

☑ Show GC contents

Sequence editor, enzyme settings.



☐ Enterokinase
☐ Factor Xa
☐ Formic acid
☐ Glutamyl
☐ GranzymeB
☐ Hydroxylamine
☐ Iodosobenzoic acid
☐ LysC
☐ NTCB
☐ Pepsin pH1.3 (1)
☐ Pepsin pH1.3 (2)
☐ Pepsin pH>2 (1)
☐ Pepsin pH>2 (2)
☐ Proline-endopeptidase
☐ Proteinase K
☐ Staphylococcal peptidase I
☑ TEV-Protease
☐ Thermolysin
☑ Thrombin
☑ Thrombin (2)

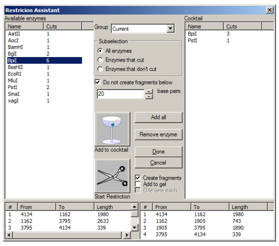
Thermolysin
!DE|AFILMV

The sequence editor, protease tab.

7.2.22 Restriction Assistant

The Restriction Assistant can be invoked via menu ***Tools/Enzyme Assistant***, or through a click with the middle mouse button on a restriction site in the DNA map. For the latter, the selected enzyme is automatically selected in the list of "Available enzymes" (left). This list depends on the selections "Group" and "Subselection". It can be sorted by enzyme name or number of cuts by clicking on the respective column title. For a selected enzyme, the resulting fragments are shown in the lower left list.

The list on the right shows the contents of the "restriction cocktail", the enzymes already selected for cutting. The resulting fragments for these enzymes together are shown in the lower right list. The enzyme selected in the left list can be put in the cocktail via ***Add to cocktail***; all enzymes from the left list can be added at once via ***Add all***. An enzyme can be removed from the cocktail by selecting it in the right list, then via ***Remove enzyme***.



Restricion Assistant

Available enzymes

Name	Cuts
AatII	1
AocI	1
BamHI	1
BglI	2
BpiI	6
BssHII	1
EcoRI	1
MluI	1
PstI	2
SmaI	1
xagI	1

Group: Current

Subselection: ● All enzymes ○ Enzymes that cut ○ Enzymes that don't cut

☑ Do not create fragments below 20 base pairs.

☑ Create fragments
□ Add to gel
□ One lane each

Cocktail

Name	Cuts
BpiI	3
PstI	1

#	From	To	Length
1	4134	1162	1980
2	1162	3795	2633
3	3795	4134	339

#	From	To	Length
1	4134	1162	1980
2	1162	1905	743
3	1905	3795	1890
4	3795	4134	339

The Restriction Assistant.

Do not create fragments below X base pairs, when selected, limits the fragments generated to a minimum size. ***Done*** exits the restriction assistant while preserving the changes made to the cocktail, whereas ***Cancel*** will void all changes made.

Start restriction (the scissors symbol) will initiate the simulated restriction. The result of this can be influenced by several further settings:

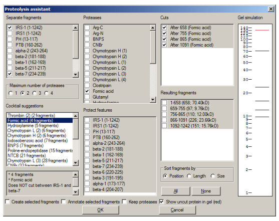
- ***Create fragments*** will generate the actual DNA sequences with their blunt/sticky ends that will result from a digestion with the cocktail. This option is pre-selected.
- ***Add to gel*** will add the fragments to a Virtual gel, together in one lane.
- ***One lane each*** will alter the above so that each enzyme gets its own lane.
- ***Partial restriction*** will add all possible fragments to a virtual gel lane, simulating a partial (incomplete) restriction. The option ***One lane each*** is not available when ***Partial restriction*** is checked.

The restriction cocktail will be preserved so you can cut another DNA with that very enzyme combination, which is useful for an upcoming ligation.

7.2.23 PCR troubleshooting dialog

This dialog can be invoked from the DNA module. It tests the primers in the module for typical problems. Note that this is ***not*** necessarily a list of errors, merely an indicator of ***potential*** problems.

7.2.24 Proteolysis assistant



Proteolysis assistant

Separate fragments
☑ IRS-1 (1-1242)
□ IRS1 (1-1242)
□ PH (13-117)
□ PTB (160-262)
□ alpha-2 (243-264)
□ beta-2 (181-188)
□ beta-1 (162-169)
□ beta-5 (211-217)
☑ beta-7 (234-239)

Maximum number of proteases
○ 1 ● 2 ○ 3 ○ 4

Cocktail suggestions
Thrombin (2) (2 fragments)
Formic acid (4 fragments)
Hydroxylamine (5 fragments)
Chymotrypsin L (2) (6 fragments)
Chymotrypsin H (2) (6 fragments)
Iodosobenzoic acid (7 fragments)
BNPS (7 fragments)
Proline-endopeptidase (15 fragments
NTCB (21 fragments)
Chymotrypsin L (3) (28 fragments)
CNBr (32 fragments)

* 4 fragments
* Formic acid
Does NOT cut between IRS-1 and beta-7

Proteases
□ Arg-C
□ Arg-N
□ BNPS
□ CNBr
□ Chymotrypsin H (1)
□ Chymotrypsin H (2)
□ Chymotrypsin L (1)
□ Chymotrypsin L (2)
□ Chymotrypsin L (3)
□ Chymotrypsin L (4)
□ Clostripain
☑ Formic acid
□ Glutamyl
□ Hydroxylamine

Protect features
□ IRS-1 (1-1242)
□ IRS1 (1-1242)
□ PH (13-117)
□ PTB (160-262)
□ alpha-2 (243-264)
□ beta-2 (181-188)
□ beta-1 (162-169)
□ beta-5 (211-217)
□ beta-7 (234-239)
□ beta-6 (220-225)
□ beta-3 (191-195)
□ alpha-1 (173-177)
□ beta-4 (204-207)

Cuts
☑ After 658 (Formic acid)
☑ After 755 (Formic acid)
☑ After 865 (Formic acid)
☑ After 1091 (Formic acid)

Resulting fragments
□ 1-658 (658; 70.40kD)
□ 659-755 (97; 9.76kD)
□ 756-865 (110; 12.00kD)
□ 866-1091 (226; 23.66kD)
□ 1092-1242 (151; 15.76kD)

Sort fragments by
● Position ○ Length ○ Size

□ Create selected fragments □ Annotate selected fragments □ Keep proteases ☑ Show uncut protein in gel (red)

The proteolysis assistant.

The ***proteolysis assistant*** can be invoked from the protein module. It can help with choosing proteases for limited proteolysis, separation of protein domains, etc. There are several list boxes which depend on each other, so it may appear confusing initially.

- First and foremost, the protein is virtually cleaved by the proteases checked in the Proteases list. The cuts from all these proteases in the protein are shown in the Cuts list. The fragments resulting from these cuts are shown in the Resulting fragments list, as well as in the Gel simulation.
- If you know that certain cuts do not take place (e.g., because the recognition motif is protected) you can deselect (uncheck) these cuts in the Cuts list.
- If you know that certain annotated features of the protein are protected from proteases, you can check them in the Protect features list.
- If you want to separate annotated features(e.g., domains) of your protein via proteases, you can select two or more features in the Separate fragments list. Depending on your selection and the state of Maximum number of proteases, the Cocktail suggestions list will be filled with proteases (or combinations of such) that will do the job. Selecting one of these "cocktails", a more detailed description will be shown below the list, and the respective proteases will be selected in the Proteases list.

The final result of all these settings is the Resulting fragments list. You can Sort fragments by cut position, fragment length, or size (weight). You can check individual fragments, All or None of them. Upon OK, several actions can be performed ***on the selected fragments***:

- Create selected fragments will create a new protein module for each selected fragment

- Annotate selected fragments will add a feature for each of the selected fragments in the current protein module
- Keep proteases will add the proteases selected in the Proteases list to the current protein module

7.2.25 Projects

A project in GENtle is a collection of sequences that belong together, even if they are in different databases. Projects can be

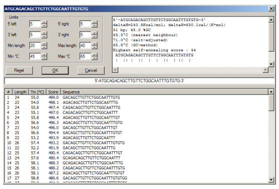
- loaded via ***File/Load Project*** or F11
- saved via ***File/Save Project*** or F12
- closed via ***File/Close Project***

Depending on the options, the last used project is automatically opened when GENtle starts.

Projects consists of a list of sequences, ***not*** the sequences themselves. If a sequence is renamed, moved or deleted, GENtle will display a warning next time a project containing that sequence is opened.

For efficient use of sequencing primers, one can create a project that contains all available sequencing primers, and then refer to that project in the Sequencing primers dialog.

7.2.26 Edit primer dialog


ATGCAGACAGCTTGTTCTGGCAATTTGTGTG

Limits

5' left 5 | 5' right 5

3' left 5 | 3' right 5

Min length 20 | Max length 40

Min °C 45 | Max °C 65

```
5'-ATGCAGACAGCTTGTTCTGGCAATTTGTGTG-3'
deltaH=240.8Kcal/mol; deltaS=630.1cal/(K*mol)
31 bp; 45.0 %GC
63.3°C (nearest neighbour)
71.0°C (salt-adjusted)
65.8°C (GC-method)
Highest self-annealing score : 44
 ATGCAGACAGCTTGTTCTGGCAATTTGTGTG
 |  || |  ||  |  |  ||  | ||  |
```

5'-ATGCAGACAGCTTGTTCTGGCAATTTGTGTG-3'

#	Length	Tm (°C)	Score	Sequence
1	24	55.0	499.0	GACAGCTTGTTCTGGCAATTTGTG
2	23	54.0	498.1	AGACAGCTTGTTCTGGCAATTTG
3	24	55.8	497.4	CAGACAGCTTGTTCTGGCAATTTG
4	23	55.1	497.0	CAGACAGCTTGTTCTGGCAATTT
5	24	55.8	496.4	AGACAGCTTGTTCTGGCAATTTGT
6	25	56.6	496.3	AGACAGCTTGTTCTGGCAATTTGTG
7	23	54.0	496.0	GACAGCTTGTTCTGGCAATTTGT
8	25	56.6	494.4	GACAGCTTGTTCTGGCAATTTGTGT
9	22	53.2	493.9	AGACAGCTTGTTCTGGCAATTT
10	26	57.4	493.2	GACAGCTTGTTCTGGCAATTTGTGTG
11	22	52.2	491.9	GACAGCTTGTTCTGGCAATTTG
12	25	57.4	490.4	CAGACAGCTTGTTCTGGCAATTTGT
13	24	57.6	490.4	GCAGACAGCTTGTTCTGGCAATTT
14	25	58.1	488.3	GCAGACAGCTTGTTCTGGCAATTTG
15	26	58.1	488.2	CAGACAGCTTGTTCTGGCAATTTGTG
16	26	58.1	487.2	AGACAGCTTGTTCTGGCAATTTGTGT
17	27	58.8	486.6	GACAGCTTGTTCTGGCAATTTGTGTGG

Edit primer dialog.

This dialog assists in optimizing a primer. For that reason, many variants of the primer are generated and can be examined. The center line of the dialog shows the current variant of the primer; details of that variant are shown in the upper right box. ***OK*** will end the dialog, committing that variant to the PCR module. ***Cancel*** will end the dialog and ***not*** change the PCR module. ***Reset*** will return the primer in the dialog to the variant the dialog was originally started with.

The list in the lower half of the dialog contains an automatically generated list of variants of the current primer, sorted by an arbitrary score. The "region" of variants can be influenced by multiple settings in the upper left quarter of the dialog. Available settings include:

- The variation of the 5'-end of the primer to the right and to the left.
- The variation of the 3'-end of the primer to the right and to the left.
- The minimum and maximum length of the primer.
- The minimum and maximum melting temperature of the primer.

Any change of these settings will trigger a recalculation of possible variations. These variations are then evaluated and shown in the list in the lower half of the dialog. Double-clicking one of the variations will change the current variation in the center line, and the properties display in the upper right quarter of the dialog.

7.2.26.1 *Properties display*

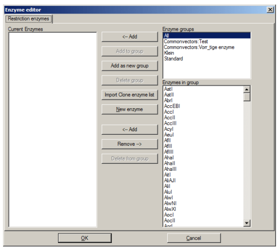
This will display:

- The primer sequence in 5'→3' orientation
- The ΔH and ΔS values
- The length and GC contents of the primer
- The melting temperature, calculated according to the Nearest Neighbour method (usually best results, but only for longer primers)
- The melting temperature, calculated according to the salt-adjusted method (mediocre results)
- The melting temperature, calculated according to the GC method (simplistic)
- The highest self-annealing score (arbitrary) and the display of that annealing
- ***Caveat :*** Calculating primer melting temperatures is tricky. If one of the three methods gives a totally different result than the other two, ignore it. Also, the melting temperature is only calculated for the 3'-end of the primer that anneals with the sequence!

7.2.27 Printing

Sequence and DNA maps can be printed via the respective context menus or the ***File*** menu. For DNA sequences, a report can be printed via ***File/Print report***. It contains the DNA map and a list of the features annotated in the sequence. This can be useful for a detailed overview of the sequence where the sequence itself is not required.

7.2.28 Enzyme management



The enzyme management dialog

The enzyme editor for enzyme management, both globally and per DNA sequence, is divided into three lists:

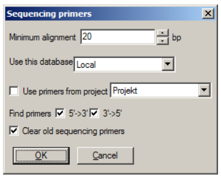
- A list of enzyme groups (top right)
- A list of enzymes in that group (bottom right)
- A list of current/temporary enzymes (left)

Enzymes can be copied into/removed from the left list through the *Add* and *Remove* buttons. Enzymes can be deleted from a group (except All) via Delete from group, or added via New enzyme. A double click on an enzyme name in either list shows an enzyme properties dialog.

Enzymes from the left list can be added to a new or existing group via the respective buttons. All enzymes from a group can be added to the left list, and a group can be deleted.

You can share enzyme groups with other GENtle users on your intranet via a commonly shared database. Create an enzyme group as "Database name:Enzyme group name", and it will be available to everyone using that database on the next start of GENtle. Note that when creating the enzyme group, use the name of the database as it appears in your local installation (maybe "Shared", "Shared0", "Shared (2)" etc.).

7.2.29 Sequencing Primers



Minimum alignment 20 bp
Use this database Local
□ Use primers from project Projekt
Find primers ☑ 5'->3' ☑ 3'->5'
☑ Clear old sequencing primers

The sequencing primers dialog can add possible sequencing primers as features to a DNA sequence. What primers to add can be specified:

- The minimum alignment (3') of a primer to the sequence. This means exact annealing!
- The database to search for primers. All primers from that database will be considered.

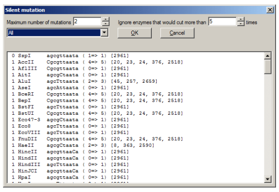
- Alternatively, use all primers that are part of a project in that database. That way, a range of primers across databases can be specified in a project and be considered as sequencing primers here.
- Primers that run in 5'→3' or 3'→5' direction.

You can also have the dialog remove old sequencing primers from the sequence. This can also be done manually through ***Edit/Remove sequencing primers*** in the DNA module. Note: Sequencing primers, if not removed, will be saved as features together with the sequence; they can still be removed lated, though.

Sequencing primers will display as yellow features, where the shade of yellow depends on their direction. The actual sequencing primer feature is only as long as the 3' annealing of the primer, so the primer might actually be longer than the feature towards the 5' end. For details, see the feature description, which contains the original primer sequence, among other data.

7.2.30 Silent Mutagenesis

This dialog can find restriction enzymes that cut in a marked DNA sequence (context menu ***Selection/Show enzymes that cut here*** in the DNA module). It can also find alternate versions of the DNA which will translate into the same amino acid sequence, but contains a new restriction site (silent mutation). A chosen enzyme/mutation will appear in the sequence (DNA or primer, respectively) upon OK.



Maximum number of mutations 2

Ignore enzymes that would cut more than 5 times

All

```
 0 SspI     agcgttaata ( 1=> 1) [2961]
 1 AccII    Cgcgttaata ( 4=> 5) [20, 23, 24, 376, 2518]
 1 AflIII   Cgcgttaata ( 0=> 1) [2961]
 1 AitI     agcgCtaata ( 0=> 1) [2961]
 1 AluI     agcTttaata ( 2=> 3) [45, 257, 2659]
 1 AseI     agcAttaata ( 0=> 1) [2961]
 1 BceRI    Cgcgttaata ( 4=> 5) [20, 23, 24, 376, 2518]
 1 BepI     Cgcgttaata ( 4=> 5) [20, 23, 24, 376, 2518]
 1 BstFI    agcTttaata ( 0=> 1) [2961]
 1 BstUI    Cgcgttaata ( 4=> 5) [20, 23, 24, 376, 2518]
 1 Eco47-3  agcgCtaata ( 0=> 1) [2961]
 1 Eco8     agcTttaata ( 0=> 1) [2961]
 1 EcoVIII  agcTttaata ( 0=> 1) [2961]
 1 FnuDII   Cgcgttaata ( 4=> 5) [20, 23, 24, 376, 2518]
 1 HaeII    agcgCtaata ( 2=> 3) [8, 363, 2590]
 1 HincII   agcgttaaCa ( 0=> 1) [2961]
 1 HindII   agcgttaaCa ( 0=> 1) [2961]
 1 HindIII  agcTttaata ( 0=> 1) [2961]
 1 HinJCI   agcgttaaCa ( 0=> 1) [2961]
 1 HpaI     agcgttaaCa ( 0=> 1) [2961]
```

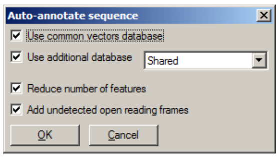
The silent mutation dialog

The results can be changed by

- changing which enzyme group to search
- limiting the number of times an enzyme may cut in the whole sequence
- limiting the number of mutations needed for a restriction site to manifest (PCR module only)

7.2.31 Automatic annotation

The ***automatic annotation*** feature can search a database of standard vectors (included with the GENtle package), and (optionally) a user-generated database, for feature sequences that are found in the currently opened DNA sequence. Recognized features are then annotated in the current sequence.



☑ Use common vectors database

☑ Use additional database Shared

☑ Reduce number of features

☑ Add undetected open reading frames

The automatic annotation dialog.

Invoked through ***<u>Edit/Auto-annotate sequence</u>*** or F9, a dialog opens, offering various settings:

- Whether or not to search the common vectors database
- Whether or not to use a user-generated database (and, if so, which one)
- Whether or not to reduce the number of generated features (recommended; otherwise, a lot of features are annotated)
- Whether or not to add unrecognized open reading frames as features

7.2.32 Sequencing Assistant Dialog

The Sequencing Assistant can ease the comparison of sequencer data with an existing sequence. Given one or two sequencer data sets and a sequence to compare them to, the assistant will invert/complement the sequencer data if necessary, then align the sequences with the appropriate algorithm.

7.3 Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Professor Dr. Helmut W. Klein betreut worden.

Köln, den

Magnus Manske

7.4 Partial Publications

- Poster presentation at the BioPerspectives congress in Wiesbaden, Germany (May 2004)
- Poster presentation at the BioRiver congress in Aachen, Germany (March 2005)
- Poster presentation at the BioPerspectives congress in Wiesbaden, Germany (May 2005)
- "GENtle - an integration and visualization of multiple algorithms", submitted September 2006

7.5 Curriculum Vitae

Magnus Manske
Elisabeth-Breuerstraße 11
51065 Cologne, Germany

Date of Birth	May 24, 1974
Place of Birth	Cologne, Germany
Nationality	German
Marital Status	single

School Education	(1980-1984)	Primary School, Cologne	
	(1984-1993)	Secondary School, Cologne	
	(1993)	Allgemeine Hochschulreife	
Study	(1993-2003)	University of Cologne	
		Biology (diploma)	
Internship	(1995-1996)	Kansas City, Kansas, USA	
Diploma	(2003)	Biology (University of Cologne)	
		Major subject	Biochemistry
		Minor Subjects	Genetics, Computer Science
Doctoral Thesis	(2003-2006)	University of Cologne	
		Group of Prof. Dr. H. W. Klein	

Lebenslauf

Magnus Manske
Elisabeth-Breuerstraße 11
51065 Köln

Geburtsdatum	24. Mai 1974
Geburtsort	Köln
Familienstand	ledig
Schulausbildung	(1980-1984) Grundschule Köln (1984-1993) Gymnasium Köln
Schulabschluss	Allgemeine Hochschulreife (1993)
Studium	(1993-2003) Universität zu Köln Diplomstudiengang Biologie
Auslandspraktikum	1995-1996 Kansas City, Kansas, USA
Diplom	Diplom-Biologe (Universität zu Köln, Januar 2003) Hauptfach Biochemie, Nebenfächer Genetik und Informatik
Doktorarbeit	2003 – 2006 Universität zu Köln Betreuer : Prof. Dr. Helmut W. Klein