Genlight: An Interactive System for High-throughput Sequence Analysis and Comparative Genomics

Michael Beckstette*† Alexander Sczyrba* Paul M. Selzer‡

Introduction

We present *Genlight*, a versatile software system to solve a wide spectrum of tasks in genome scale sequence analysis, with a special focus on differential comparative genome analysis [BMM+04]. The system is designed for (i) the discovery of potential new drug targets by comparative genome analysis, (ii) automatic genomic scale analyses in reasonable time, without the need for specialized hardware or large cluster systems, (iii) the integration of various bioinformatics sequence analysis methods and the storage of their results in a structured reusable and queryable way and (iv) dynamic result presentation and visualization through an easy to use, but flexible interface. The reuseability of results is a central concept in our system, achieved by a set oriented data model. The *Genlight* system is multi-user capable, suited for high-throughput analysis of biomolecular data. It combines the advantages of an object relational database management system with a distributed client server approach for large scale compute tasks and provides a powerful user interface. We present two case studies, in which *Genlight* was used to detect a new gene family in maize and to analyze EST clusters of the african claw frog *Xenopus laevis*.

System architecture and functionality

The *Genlight* system consists of four major parts:(i) a web based user interface, (ii) the *Genlight* server scheduling and distributing various bioinformatics analysis tasks to (iii) the client components which carry out these tasks in an asynchronous way, and (iv) a database component for storing, modifying, and accessing data. The underlying relational database system allows easy access to the generated data for the built-in software tools as well as for external applications.

The comparison of whole genomes or proteomes and their use as query sets for searches in large databases like Genbank or Swiss-Prot [BBA+03] is challenging and time consuming. To handle such comparison tasks in an interactive system, the individual comparison calculations have to be performed asynchronously. Therefore, the *Genlight* server com-

*Technische Fakultät, Universität Bielefeld, Postfach 100 131, D-33501 Bielefeld, Germany
†Corresponding author, Email: mbeckste@techfak.uni-bielefeld.de
‡Akzo Nobel, Intervet Innovation GmbH, BioChemInformatics, Zur Propstei, D-55270 Schwabenheim, Germany

ponent contains a queuing mechanism for all analysis tasks. To process queued entries, the system has its own scheduling component, that allows a parallel, distributed execution of comparison jobs and can form a virtual cluster system of regular workstations for high throughput analysis tasks. The two major strengths of this approach are the complete integration into one system and a high level of robustness of the system. The latter is achieved by methods to insure data integrity, like a backlog technique and connection supervision, during distributed execution. Compute nodes can be added to and deleted from the virtual cluster system at any time, by starting or stopping the *Genlight* client component on a workstation. Due to the flexibility of the virtual cluster it is possible to temporarily include or exclude certain computers at any time. This distributed computing approach allows comparisons of complete genomes or proteomes in short time intervals. Even time consuming and CPU intensive tasks like Hidden Markov Model based approaches can be processed in reasonable running times.

The structured storage, ensuring reuseability of generated results, is a critical point for the protocol based step by step modeling of complex experiments and workflows, often neglected in bioinformatics applications. In *Genlight* the re-use of derived results is a central concept, anchored in the basic system design. It is achieved by a set oriented data model with only two basic data object types: *Seq-set*s and *Hit-set*s. A *Seq-set* is a collection of sequences of one type, either nucleotide or protein. A *Hit-set* is a set of sequence pairs, defined by a comparison operation between two *Seq-set*s and its user defined parameterization, e.g. the set of all sequence pairs detected by a homology search between two seq-sets. *Genlight* supports various operations that can be applied to *Hit-set*s or *Seq-set*s; each operation resulting in a new *Seq-set* or *Hit-set*. A number of *Hit-set* filters are predefined, for instance selecting those pairs, where the aligned region of the query sequence covers the complete target sequence (see *X.laevis* case study for an example). Additionally, users can add custom made filters. These filters are applied to *Hit-set*s generating new *Hit-set*s. Sequence filters generate new *Seq-set*s and extraction operations convert a *Hit-set* to a new *Seq-set* according to the specified criterias. This procedure follows the software engineering concept of *compositionality* and allows an interactive step by step modelling of complex workflows as schematically shown in Figure 1.

Using a combination of comparison, filter, and extraction operations, several proteomes, say A, B, and C, can be easily screened for proteins common to the proteom sets A and B but nonexistent in proteome set C. Moreover, all possible intersections of A, B, and C can be calculated. Evidence of proteins with similar function can be defined by combinations of several homology search results (e.g. unidirectional or bidirectional best hits), even generated by different homology search methods.
An integrated project management, providing fundamental access control features, allows to store seq-sets and hit-sets on a per-user basis. Frequently used seq-sets and hit-sets, like major public sequence databases such as Genbank, SpTrembl and genomes or proteomes of model organisms and their analysis results, can be made available system-wide, to save resources and avoid redundancy. The administrative features are completed by a quota system, that allows to restrict resources on a per-user and per-method basis.

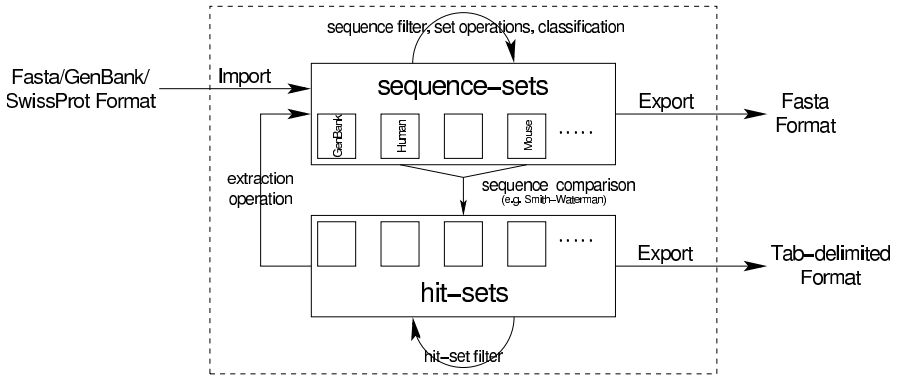

Figure 1: The set-oriented concept: Basic data structures and their compositionality

Integrated sequence analysis methods and databases

Genlight can handle nucleotide as well as protein sequences. Almost all algorithms of the BLAST [AMS+97], and FASTA [Pe94] family as well as the traditional Smith-Waterman algorithm [SW81] are integrated.
For the discovery of conserved sequence motifs, we integrated the following motif databases: the protein family databases Pfam [BLD+00] and Tigrfam [HSW03], the conserved domain database CDD [MBPS+02] and the SMART database [LCS+04] with their specific search routines *hmmpfam* [DEK98] and *reverse position specific blast*. For the functional/structural classification of sequences we integrated the COG [TNG+01], its eukaryotic complement KOG [TFJ+03] and the SCOP [AHB+04] database. Moreover, Gene Ontologies (GOs) can be assigned by the system, inferred from the respective assignment of the integrated databases. In addition, any sequence databases available in Genbank, Swissprot or FASTA format can be imported. These databases are treated like normal *Seq-set*s and can be made available as a system-wide resource by the *Genlight* administrator. The modular architecture of the system allows the straightforward integration of new analysis methods.

The *Genlight* user interface

Our system is completely web-accessable and makes use of dynamic data visualization with the help of the GD graphics library. Calculated results are presented in graphical as well as in tabular form.

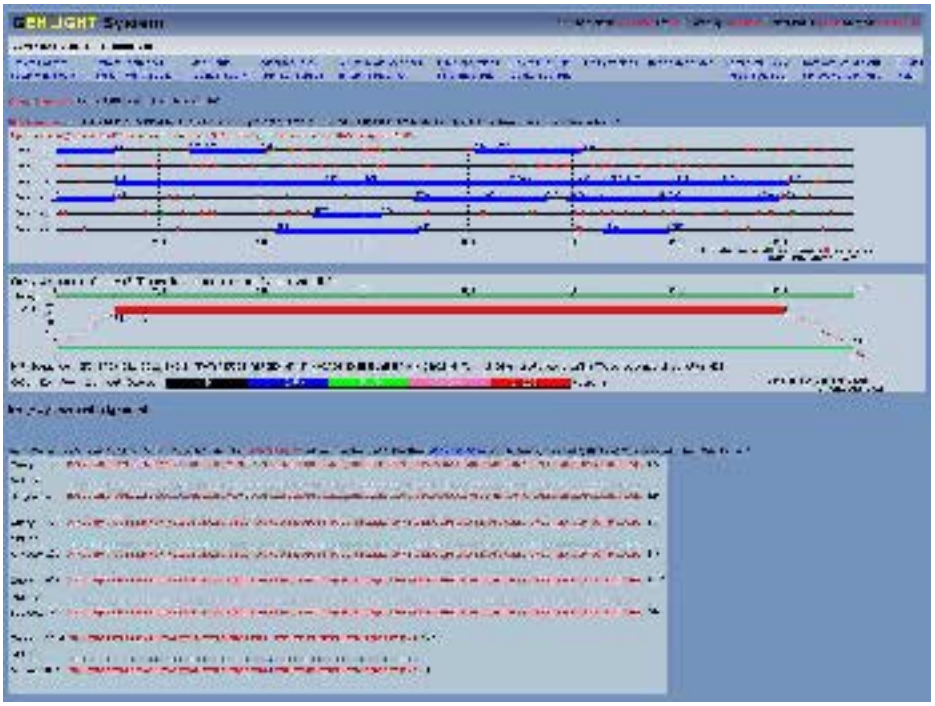

Figure 2: Screenshot showing graphical and colored textual representation of a FASTY alignment. The upper graphic shows the open reading frames (blue boxes), the middle graphic represents the matching region in query and target sequences and the textual alignment is shown at the bottom.

Case Studies

Detection of the *Smh* gene family in maize

In [MBG+03] we detected a new gene family in maize (*Zea mays*), called *Single myb histone (Smh)* family, with the help of the *Genlight* prototype. We screened GenBank, ZmDB and Pioneer Hi-Bred (PHI) expressed sequence tag (EST) databases with *Genlight*'s built-in sequence comparison and motif search methods for the occurence of the myb-like domain of human TRF1. We identified several maize ESTs that encoded proteins with a single N-terminal myb-like domain. Together, the EST and additional cDNA library screens uncovered cDNAs from five related genes. The deduced protein sequences from five different full-length cDNAs revealed a family of small basic proteins. The cDNAs were from an uncharacterized gene family, the *Smh* gene family. Detailed sequence analysis with *Genlight* revealed a number of surprising features of *Smh* genes. The most remarkable aspect was their triple-motif structure, which has not been previously described in any system, plant, animal, fungal, or bacterial. Namely, *Smh* genes have (a) an N-terminal myb

	NR	Human	Mouse	Rat	Fly	Worm	X.laevis	X.tropicalis
HQFL	5139	2347	2337	1930	268	190	3862	660
MQFL	9576	3528	3774	3374	473	357	4967	796
LQFL	14094	6467	6701	6341	2249	1918	5701	1241

Table 1: Number of *X. laevis* contigs with full length FASTY hits in the non-redundant protein database (NCBI), five model organisms, and available *X. laevis* and *X. tropicalis* proteins, determined by FASTY. Lower quality categories include sequences from higher, more stringent categories.

like or SANT domain of the homeodomain-like superfamily of 3-helical-bundle-fold proteins, (b) a central region with homology to the globular domain of linker histones H1/H15, and (c) a strong prediction signature for a coiled-coil domain near the C-terminus. Additional large scale database searches with *Genlight* revealed that *Smh*-type genes are plant specific and include a gene family in Arabidopsis and one gene of parsley (*Petroselinum crispum*). Various wet-lab experiments with a chosen member of the *Smh* family showed the ability to bind telomeric DNA repeats in vitro.

Analysis of *Xenopus laevis* EST clusters

Clustering EST sequences is a widely used method for analyzing the transcriptome of a genome. Especially in organisms where the genome sequence is not (yet) sequenced, the EST data is a valuable source of information. *Genlight* was used for extensive analysis of 31,353 contig and 40,877 singleton sequences resulting from clustering and assembly of 350,468 *Xenopus laevis* ESTs. The analysis focused on the identification of full length contigs, representing potential new genes from *X.laevis* (manuscript in preparation).

Sequences were subject to FASTY homology searches vs. the non-redundant protein database (NR) from NCBI, the proteomes of five major model organisms (*H. sapiens*, *M. musculus*, *R. norvegicus*, *C. elegans*, *D. melanogaster*), *X. laevis* and *X. tropicalis*.

For full length contig identification different *Hit-set* filters were defined and applied to the *Hit-set*s of the model organisms. FASTY hits were categorized into three classes: (1) High Quality Full Length (HQFL) hits were defined as matches covering 100% of the sequence of a known protein. Additionally, the matched protein sequence had to start with a methionine and the matching region of the contig had to start with the start codon ATG and the hit had to end at a STOP codon. (2) Medium Quality Full Length (MQFL) hits were defined as matches capable of covering 100% of the matched protein sequence with no additional constraints. (3) Low Quality Full Length (LQFL) hits were matches that covered the protein over almost its full length, allowing the match to start or end up to 10 amino acids after or before the start or end of the protein respectively. Table 1 shows the numbers of full length sequences matching proteins for each model organism.

For a functional classification of the clustered *X.laevis* data set, a non-redundant sequence set was built. In each cluster, a single contig was selected, resulting in 26,187 sequences.

The non-redundant data set was then classified based on homology to known proteins from the KOG database (BLASTX 1.0e-5 E-value cutoff, best hit filter). 17,624 sequences (67.3%) had a hit against the KOG database and could be assigned a functional category.

Acknowledgement

Financial support for the development of *Genlight* was provided by the Department for BioChemInformatics of Intervet Innovation GmbH, Schwabenheim, Germany. We thank Dirk Evers for critically reading the manuscript.

Availability

The *Genlight* system is publicly available for non-commercial use at http://piranha.techfak.uni-bielefeld.de.

References

[AHB+04] Andreeva, A., Howorth, D., Brenner, S., Hubbard, T., Chothia, C., and Murzin, A.: SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32**:D226–D229. 2004.

[AMS+97] Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.: Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **25**(17):3389–3402. 1997.

[BBA+03] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M., Michoud, K., Donovan, C., Phan, I., Pibout, S., and Schneider, M.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**:365–370. 2003.

[BLD+00] Bateman, A., Lachlan, C., Durbin, R., Finn, R., Hollich, V., Griffith-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E., Studholme, D., Yeats, C., and Eddy, S.: The Pfam Protein Families Database. *Nucleic Acids Res.* **28**:263–266. 2000.

[BMM+04] Beckstette, M., Mailänder, J. T., Marhöfer, R. J., Sczyrba, A., Ohlebusch, E., Giegerich, R., and Selzer, P. M.: Genlight: Interactive high-throughput sequence analysis and comparative genomics. *submitted*. 2004.

[DEK98] Durbin, R., Eddy, S., and Krogh, A.: *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge University Press. New York. 1998.

[HSW03] Haft, D., Selengut, J., and White, O.: The TIGRFAMs database of protein famalies. *Nucleic Acids Res.* **31**:371–373. 2003.

[LCS+04] Letunic, I., Copley, R., Schmidt, S., Cicacarelli, F., Doerks, T., Schultz, J., Ponting, C., and Bork, P.: SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* **32**:D142–D144. 2004.

[MBG+03] Marian, C., Bordoli, S., Goltz, M., Santarella, R., Jackson, L., Danilevskaya, O., Beckstette, M., Meeley, R., and Bass, H.: The Maize *Single myb histone* 1 Gene, *Smh1*, Belongs to a Novel Gene Family and Encodes a Protein That Binds Telomere DNA Repeats in Vitro. *Plant Physiology*. **133**:1336–1350. 2003.

[MBPS+02] Marchler-Bauer, A., Panchenko, A., Shoemaker, B., Thiessen, P., Geer, L., and Bryant, H.: CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**(1):281–283. 2002.

[Pe94] Pearson, W. R.: Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.* **24**:307–331. 1994.

[SW81] Smith, T. and Waterman, M.: Identification of common molecular subsequences. *Journal of Molecular Biology*. **47**:195–197. 1981.

[TFJ+03] Tatusov, R., Fedora, N., Jackson, J., Jakobs, A., Kiryutin, B., Koonin, E., Krylov, D., Mazumder, R., Mekhedov, S., Nikolskaya, A., Rao, B., Smirnov, S., Sverdlov, A., Vasudevan, S., Wolf, Y., Yin, J., and Natale, D.: The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. **11**(4):41. 2003.

[TNG+01] Tatusov, R., Natale, D., Garkavtsev, I., Tatusova, T., Shankavaram, U., Rao, B., Kiryutin, B., Galperin, M., Federova, N., and Koonin, E.: The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**(1):22–28. 2001.