BioSmalltalk

Hernán Morales Durand - ESUG 2023 @ Lyon, France

BioSmalltalk is...

... a library for Bioinformatics

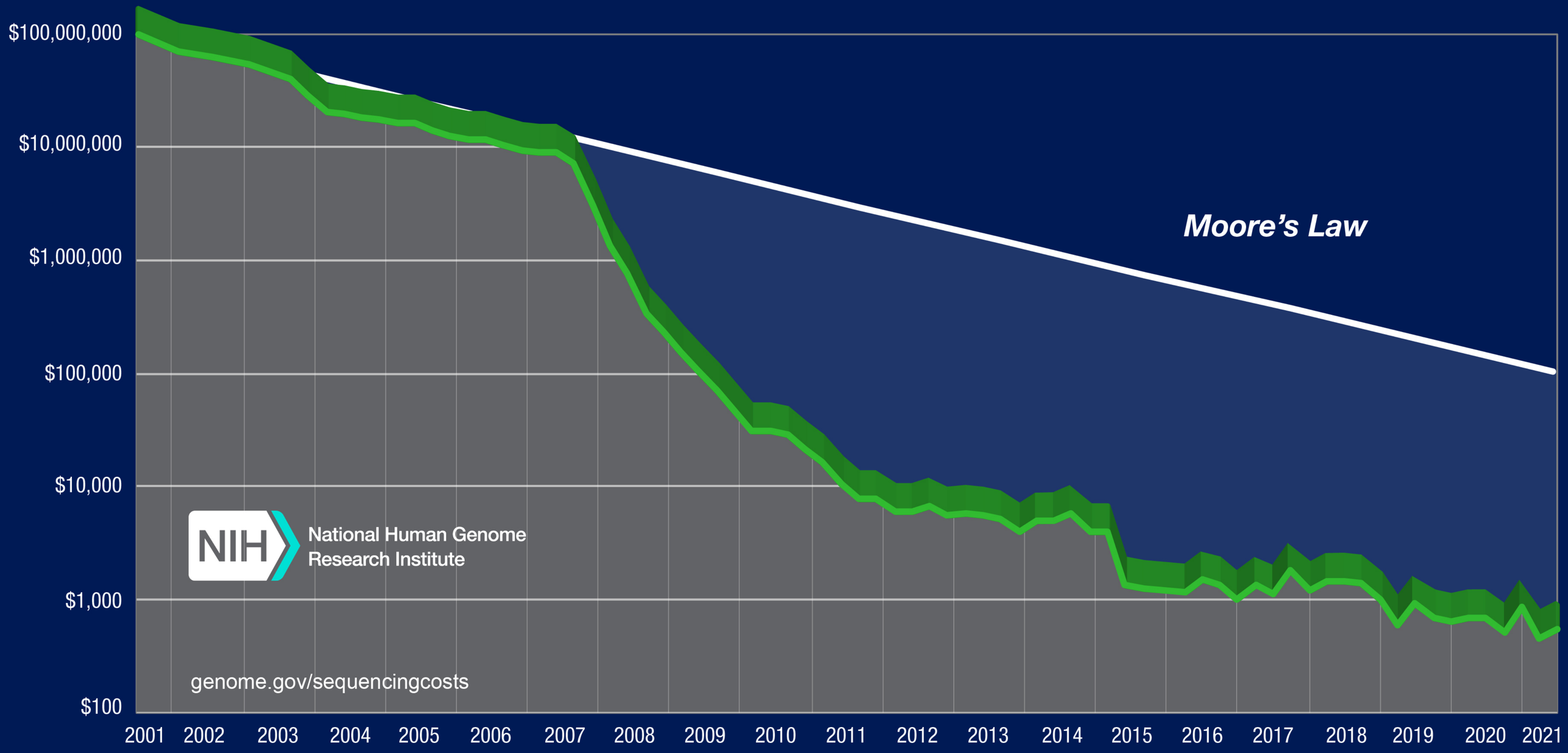
...implemented in Pharo 

...part of the Open Bioinformatics Foundation (OBF)

...not intended to be a replacement of +30000 awesome bioinformatics tools (but it could save some time).

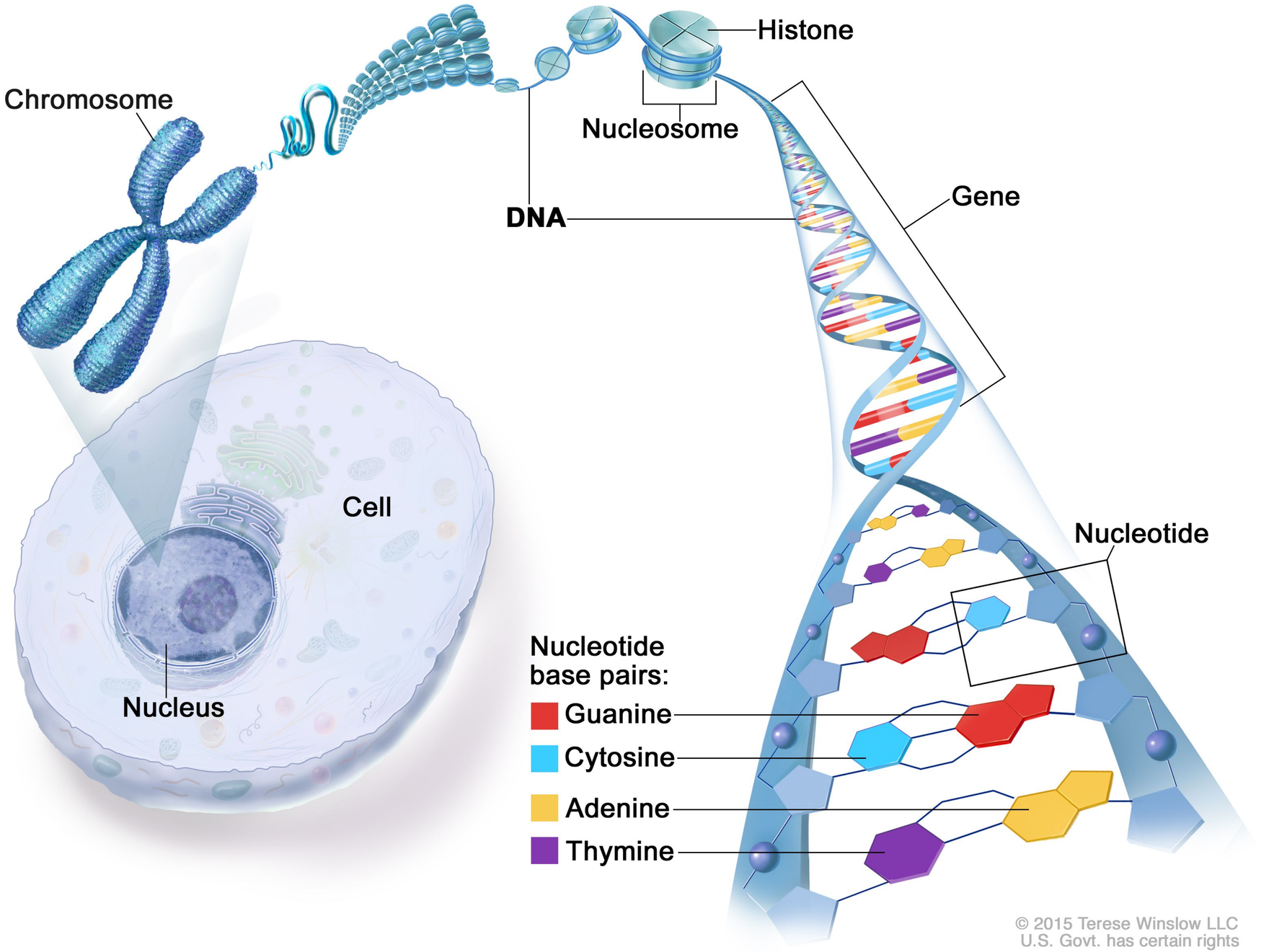
What is Bioinformatics?

Cost per Human Genome

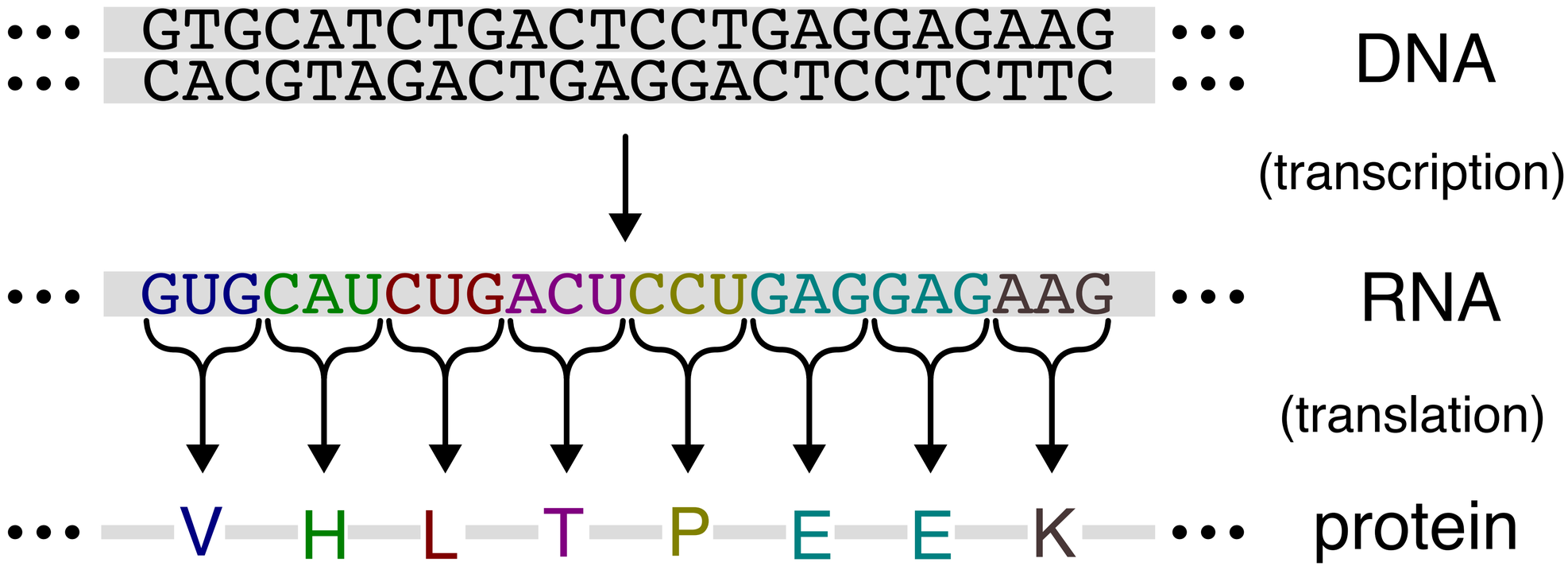


https://www.genome.gov/sites/default/files/media/files/2021-11/Sequencing_Cost_Data_Table_Aug2021.xls

DNA Structure



BioSmalltalk: Basic operations



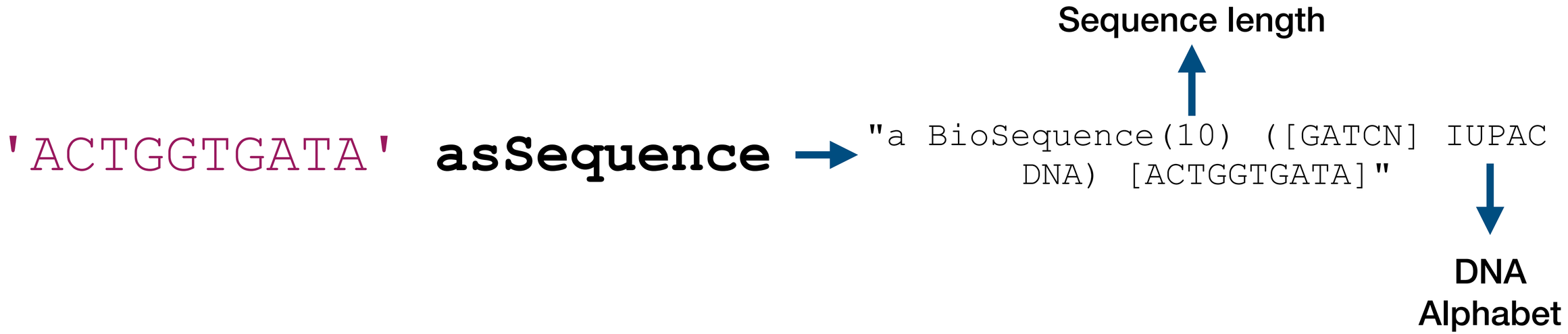
Basic operations with biological sequences

BioSmalltalk: Basic operations

```
'ACTGGTGATA' asSequence
```

```
"a BioSequence(10) ([GATCN] IUPAC DNA) [ACTGGTGATA]"
```

BioSmalltalk: Basic operations



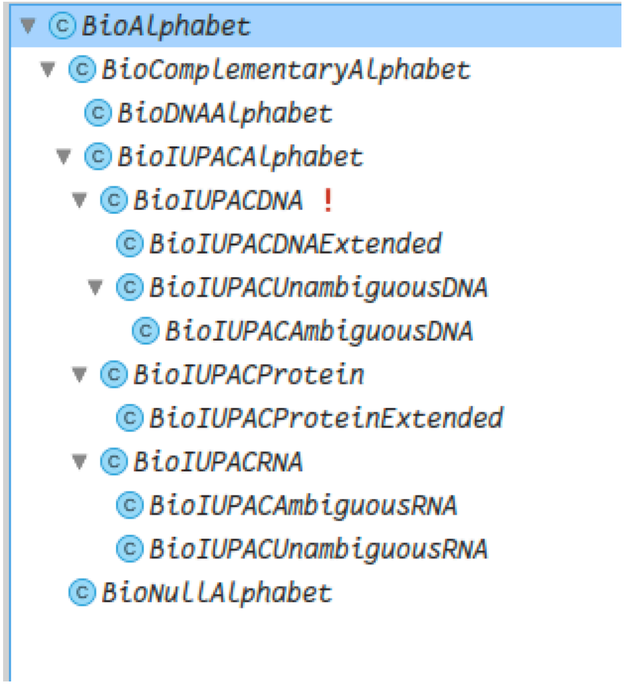
BioSmalltalk: Basic operations

```
'ACTGGTGATA' asSequence
```

→ `"a BioSequence(10) ([GATCN] IUPAC DNA) [ACTGGTGATA]"`

Sequence length

DNA Alphabet



BioSmalltalk: Basic operations

```
'ACTGGTGATA' asSequence
```

```
"a BioSequence(10) ([GATCN] IUPAC
DNA) [ACTGGTGATA]"
```

transcribe →

```
"a BioSequence(10) ([GAUC] IUPAC ->
RNA -> Unambiguous) [ACUGGUGAUA]"
```

RNA Alphabet

BioSmalltalk: Basic operations

```
'ACTGGTGATA' asSequence
```

transcribe

backTranscribe

```
"a BioSequence(10) ([GATCN] IUPAC
DNA) [ACTGGTGATA]"
```

```
"a BioSequence(10) ([GAUC] IUPAC ->
RNA -> Unambiguous) [ACUGGUGAUA]"
```

BioSmalltalk: Basic operations

```
'ACTGGTGATA' asSequence
```

```
"a BioSequence(10) ([GATCN] IUPAC DNA) [ACTGGTGATA]"
```

```
complement
```

```
"a BioSequence(10) ([GATCN] IUPAC DNA) [TGACCACTAT]"
```

BioSmalltalk: Basic operations

```
'ACTGGTGATA' asSequence
```

```
"a BioSequence(10) ([GATCN] IUPAC DNA) [ACTGGTGATA]"
```

reverseComplement

```
"a BioSequence(10) ([GATCN] IUPAC DNA) [TATCACCAGT]"
```

BioSmalltalk: Basic operations

```
'ACTGGTGATA' asSequence
```

translate

```
"a BioSequence(10) ([GATCN] IUPAC DNA) [ACTGGTGATA]"
```

```
"a BioSequence(3) ([ACDEFGHIKLMNPQRSTVWY] IUPAC -> Protein) [TGD]"
```

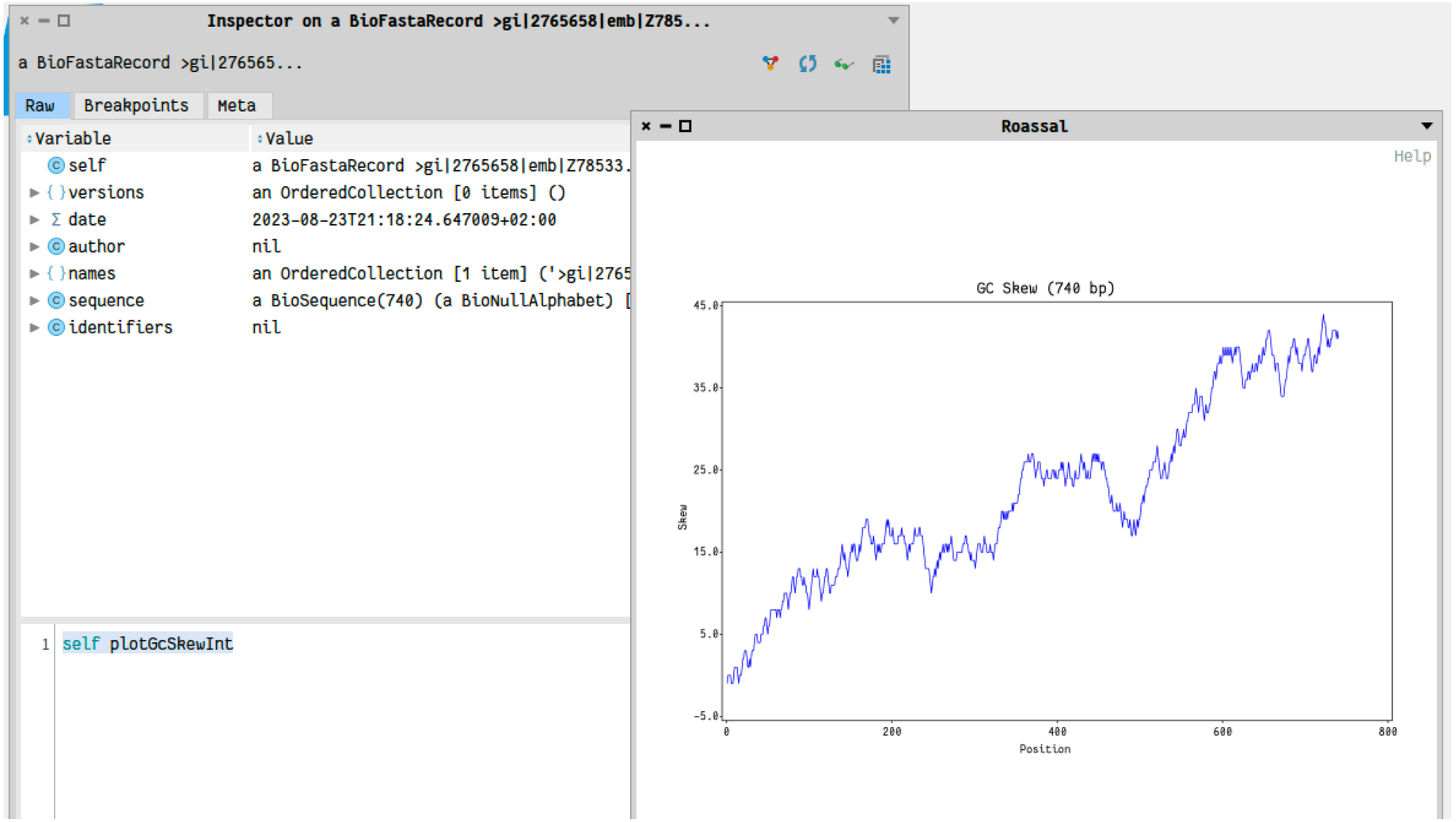
Sequence Utilities

BioSmalltalk: Sequence statistics

```
'ACTGGTGATA' asSequence gcContent → "40s0"
```

BioSmalltalk: Sequence statistics

GC Skew plot



BioSmalltalk: Sequence statistics

```
'ACTGGTGATA' asSequence gcContent

'ACTGGTGATA' asSequence molecularWeightNonDegen
```

```
"3146.0499999999997"
```

BioSmalltalk: Sequence statistics

```
'ACTGGTGATA' asSequence gcContent

'ACTGGTGATA' asSequence molecularWeightNonDegen

'ACTGGTGATA' asSequence lcc
```

```
"an OrderedCollection(-1.494867642699313
-0.16609640474436815 -1.494867642699313
                -1.494867642699313)"
```

BioSmalltalk: Sequence statistics

```
'ACTGGTGATA' asSequence gcContent

'ACTGGTGATA' asSequence molecularWeightNonDegen

'ACTGGTGATA' asSequence lcc

'ACTGGTGATA' asSequence occurrencesOfLetters
```

```
"a Dictionary($A->3
$C->1 $G->3 $T->3 )"
```

BioSmalltalk: Sequence utilities

```
(BioSequence newAmbiguousDNA:'AHT') disambiguate
```

```
"#('AAT' 'ATT' 'ACT')"
```

BioSmalltalk: Sequence utilities

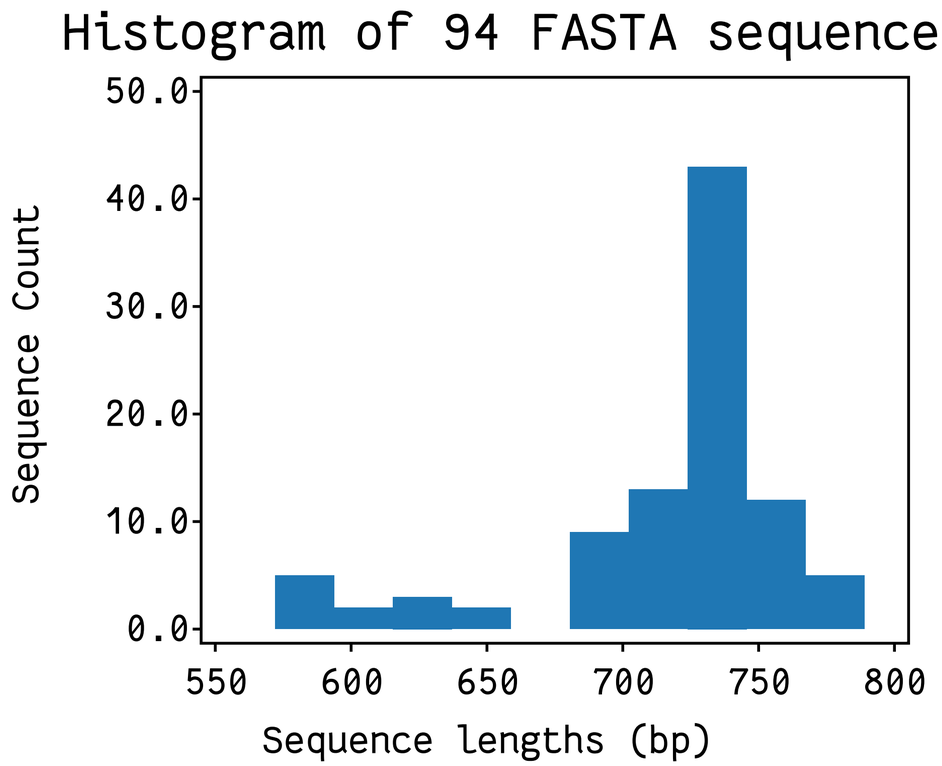
```
(BioSequence newAmbiguousDNA:'AHT') disambiguate

'ACGTACGTACGT' asSequence kmersCount:'CG'
'ACGTACGTACGT' asSequence longestConsecutive: $A

'ACTGGTGATA' asSequence crc32.
'ACTGGTGATA' asSequence gcg.
'ACTGGTGATA' asSequence seguid.
```

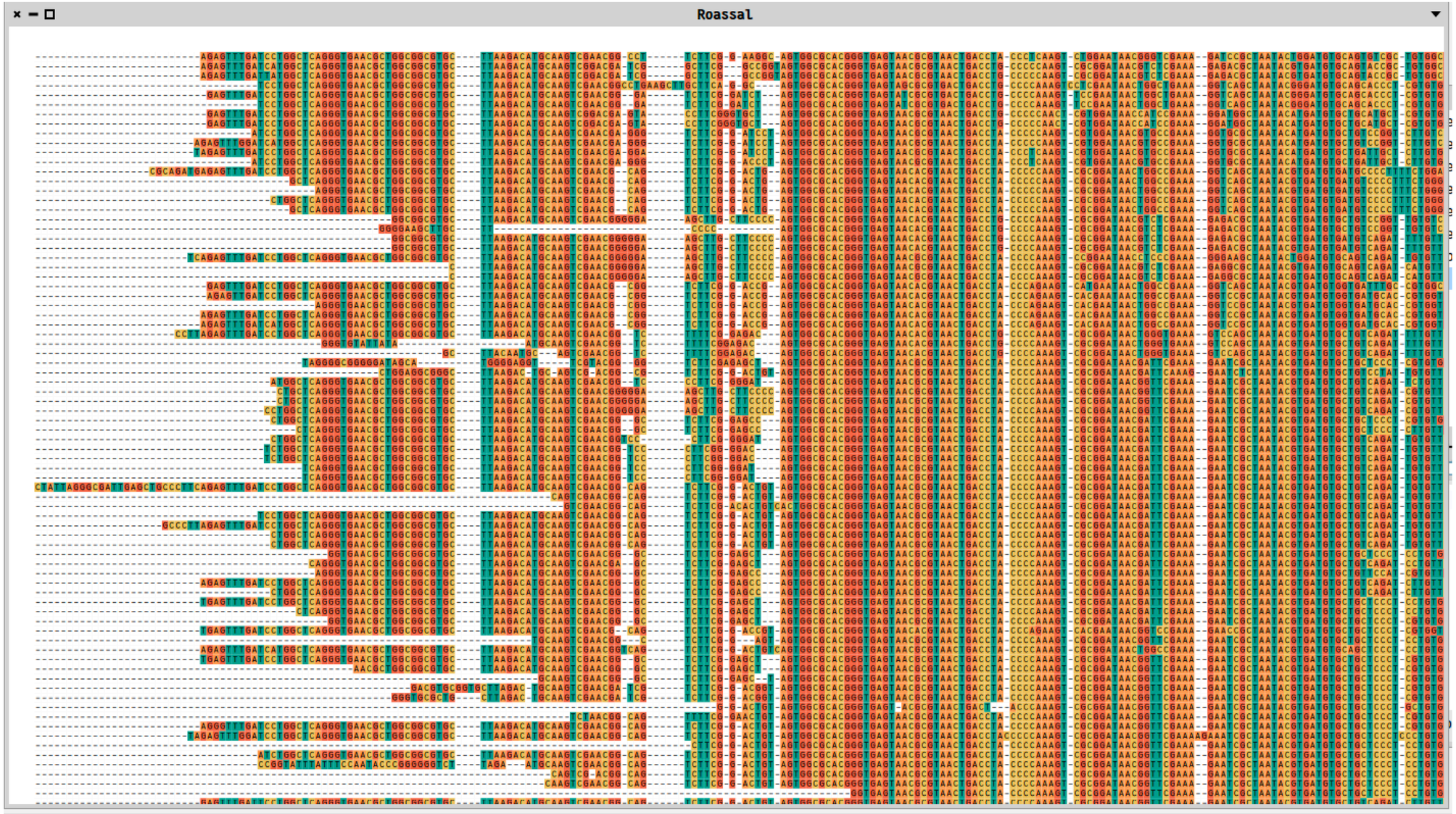
BioSmalltalk: Sequence utilities

```
(BioParser parseMultiFastaFile: 'ls_orchid.fasta') plot
```



Sequence Alignment

BioSmalltalk: Sequence alignment



BioSmalltalk: Sequence alignment

```
BioAlignment new
 addSequence: 'ACTGCTAGCTAG';
 addSequence: 'ACT-CTAGCTAG';
 addSequence: 'ACTGGTANATGG';
 addSequence: 'ACTGATTGCTGG';
 addSequence: 'ACTGCTTGATTG';
 yourself
```

BioSmalltalk: Sequence alignment

```
latestBlast := BioBlastWrapper ncbi local latest.
latestBlast nucleotide
        query: '555';
        hitListSize: 10;
        filterLowComplexity;
        expectValue: 10;
        blastn;
        blastPlainService.
```

Program	Query Type	DB Type	Comparison
blastn	Nucleotide	Nucleotide	Nucleotide-Nucleotide
blastp	Protein	Protein	Protein-Protein
tblastn	Protein	Nucleotide	Protein-Protein
blastx	Nucleotide	Protein	Protein-Protein

BioSmalltalk: Sequence alignment

```
aligner := BioMAFFTWrapper new.
aligner
    input: 'COVID-19-01.fasta';
    addOutputParameter: 'output.aln';
    execute
```

BioSmalltalk: Sequence alignment

```
aligner := ALNeedlemanWunsch new.
aligner
  align: 'AC-AATAGAC'
   with: 'ACGAATAGAT'.
```

Implementation of Needleman-Wunsh algorithm native to Pharo

https://github.com/hernanmd/needleman-wunsch

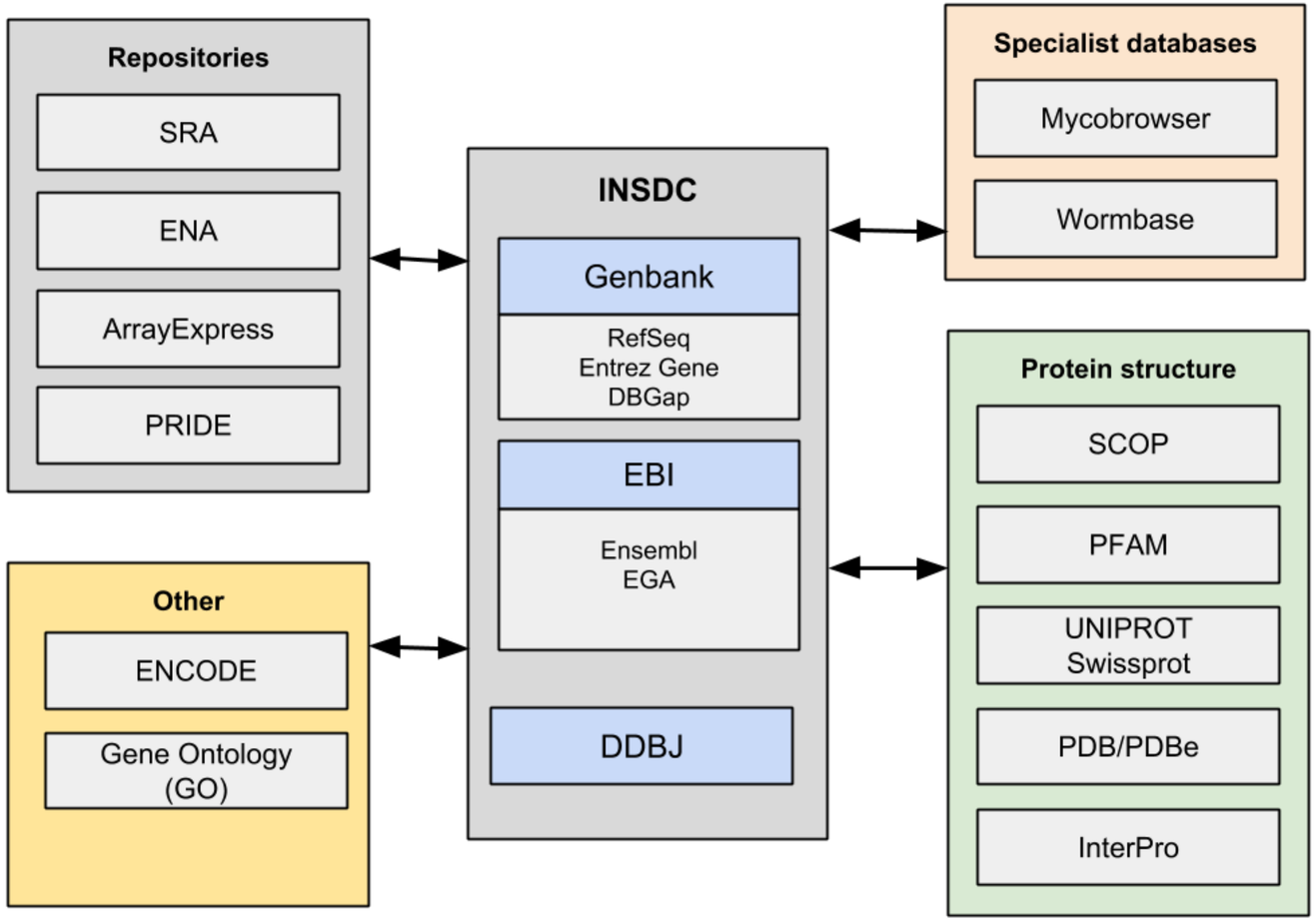
BioSmalltalk: Plotting alignment pipeline

```
outputFilename := 'COVID-19-MAFFT-2023-08-24_21-37-49.aln'.
sarsCoV2SequencesUIDs := 'seqIDs.txt' asFileReference lines.
multiFasta := BioParser parseMultiFasta: (
   BioEntrezClient new nuccore
      uids: sarsCoV2SequencesUIDs;
      setFasta;
      setModeText;
      fetch) result.
multiFastaCompleteGenomes := multiFasta
   select: [ : f | f name endsWith: 'complete genome' ].
BioMAFFTWrapper new
      auto;
      maxiterate: 1000;
      input: multiFastaCompleteGenomes;
      addOutputParameter: outputFilename;
      execute.
(BioParser parseMultiFastaAlignmentFile: outputFilename asFileReference) plot.
```

BioSmalltalk: Miscellany

- **Sequences:** Consensus, Repeats, Codon Tables, IUPAC Alphabets, Features, Records.
- **Genome downloads**
- **Wrappers:** PLINK, Cutadapt, MUSCLE, BLAST, CLUSTAL, STRUCTURE, Shapelt, HH-Suite, ACANA, AGA, samtools, etc.
- **Formatters:** FASTA, GenBank, PED, BED, MEGA, Arlequin, etc
- **Parsers:** GenBank & Entrez XML, ID's, FASTA.
- **Databases:** NCBI Entrez, REBASE.

BioSmalltalk: Databases



BioSmalltalk: Databases

```
BioEntrezClient organization listAtCategoryNamed: 'accessing public - databases' ].

"#(#gds #geo #genome #pmc #genomeprj #nlmcatalog #unigene #homologene #nucest #peptidome #journals #domains #structure #omia #omim #pubmed #biosystems #popset #cancerchromosomes #gensat #snp #books #ncbisearch #gene #pcsubstance #nuccore #protein #cdd #sra #nucgss #proteinclusters #biosample #taxonomy #unists #probe #mesh #pcassay #gap #pccompound)"
```

Applications

Animal (2018), **12:2**, pp 215–223
doi:10.1017/S1751731117001380

Evidence of positive selection towards Zebuine haplotypes in the BoLA region of Brangus cattle

D. E. Goszczynski[1†a] C. M. Corbi-Botto[1a], H. M. Durand[1], A. Rogberg-Muñoz[1,2,3], S. Munilla[2,3], P. Peral-Garcia[1], R. J. C. Cantet[2,3] and G. Giovambattista[1]

[1]*Facultad de Ciencias Veterinarias, Instituto de Genética Veterinaria (IGEVET) (UNLP-CONICET LA PLATA), La Plata, Buenos Aires, Argentina;* [2]*Departamento de Producción, Facultad de Agronomía, Universidad de Buenos Aires, Buenos Aires, Argentina;* [3]*Instituto de Investigaciones en Producción Animal (INPA) (UBA-CONICET), Ciudad Autónoma de Buenos Aires, Argentina.*

(Received 5 January 2017; Accepted 25 April 2017; First published online 14 July 2017)

*The Brangus breed was developed to combine the superior characteristics of both of its founder breeds, Angus and Brahman. It combines the high adaptability to tropical and subtropical environments, disease resistance, and overall hardiness of Zebu cattle with the reproductive potential and carcass quality of Angus. It is known that the major histocompatibility complex (MHC, also known as bovine leucocyte antigen: BoLA), located on chromosome 23, encodes several genes involved in the adaptive immune response and may be responsible for adaptation to harsh environments. The objective of this work was to evaluate whether the local breed ancestry percentages in the BoLA locus of a Brangus population diverged from the estimated genome-wide proportions and to identify signatures of positive selection in this genomic region. For this, 167 animals (100 Brangus, 45 Angus and 22 Brahman) were genotyped using a high-density single nucleotide polymorphism array. The local ancestry analysis showed that more than half of the haplotypes (55.0%) shared a Brahman origin. This value was significantly different from the global genome-wide proportion estimated by cluster analysis (34.7% Brahman), and the proportion expected by pedigree (37.5% Brahman). The analysis of selection signatures by genetic differentiation (F_{st}) and extended haplotype homozygosity-based methods (*iHS *and* Rsb*) revealed 10 and seven candidate regions, respectively. The analysis of the genes located within these candidate regions showed mainly genes involved in immune response-related pathway, while other genes and pathways were also observed (cell surface signalling pathways, membrane proteins and ion-binding proteins). Our results suggest that the BoLA region of Brangus cattle may have been enriched with Brahman haplotypes as a consequence of selection processes to promote adaptation to subtropical environments.*

Keywords: Brangus, major histocompatibility complex, selection signatures, BoLA, ancestral haplotypes

Legal Medicine 22 (2016) 54–57

Contents lists available at ScienceDirect

Legal Medicine

journal homepage: www.elsevier.com/locate/legalmed

Case Report

DNA profile of dog feces as evidence to solve a homicide

L.S. Barrientos [a,1,2], J.A. Crespi [a,1,2], A. Fameli [b], D.M. Posik [a,2], H. Morales [a,2], P. Peral García [a,2], G. Giovambattista [a,*]

[a] IGEVET – Instituto de Genética Veterinaria (UNLP-CONICET LA PLATA), Facultad de Ciencias Veterinarias, UNLP, La Plata, Buenos Aires, Argentina
[b] GECOBI – Grupo de Genética y Ecología en Conservación y Biodiversidad, Museo Argentino de Ciencias Naturales "Bernardino Rivadavia", Av. Angel Gallardo 470, C1405DJR Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 31 March 2016
Received in revised form 20 June 2016
Accepted 10 August 2016
Available online 10 August 2016

Keywords:
Forensic sciences
Non-human DNA
Dog
Mitochondrial DNA
Feces

ABSTRACT

Dog fecal samples were collected at the crime scene and from the shoes of the suspect to see whether they could be linked. DNA was genotyped using a 145 bp fragment containing a 60 bp hotspot region of the mitochondrial DNA (mtDNA) control region. Once the species origin was identified, sequences were aligned with the 23 canine haplotypes defined, showing that evidence and reference had 100% identity with haplotype 5. The frequency of haplotype 5 and the exclusion power of the reference population were 0.056 and 0.89, respectively. The forensic index showed that it was 20 times more likely that the evidence belonged to the reference dog than to some other unknown animal. The results support that the mtDNA hypervariable region 1 (HV1) is a good alternative for typing in trace or degraded casework samples when the STR panel fails, and demonstrate the utility of domestic animal samples to give additional information to solve human legal cases.

1. Introduction

Non-human DNA analysis in forensic science has seen growth in recent years. Applications range from investigations of crimes of humans to cruelty and poaching in animal/wildlife species, where DNA evidence from animals, plants, bacteria and viruses has been used in criminal investigations [1].

Animal Forensic Genetics is defined as "The application of relevant genetic techniques and theory to legal matters, for enforcement issues, concerning animal biological material" [2]. Domestic animal genetic evidence has become an important forensic tool close relationship with people, determination of the genetic profile of pets would provide a valuable forensic tool.

Canine biological materials including hair, feces and saliva can be found when contact between dogs and humans takes place. Most of the described collection, sampling, and extraction are used in medical diagnostic applications [8,9], wildlife population [10,11] and wildlife illegal traffic studies [12]. Fecal DNA is often degraded due to environmental factors and continued active deterioration by the large numbers of bacteria present with the feces. Also, feces contain many known PCR inhibitors such as bile salts [13]. As fecal samples are not commonly received in forensic laboratories, our

PLOS ONE

RESEARCH ARTICLE

Runs of homozygosity in a selected cattle population with extremely inbred bulls: Descriptive and functional analyses revealed highly variable patterns

Daniel Goszczynski[1], Antonio Molina[2], Ester Terán[1], Hernán Morales-Durand[1], Pablo Ross[3], Hao Cheng[3], Guillermo Giovambattista[1,2,3,4], Sebastián Demyda-Peyrás[1,2,3,4]*

1 IGEVET–Instituto de Genética Veterinaria "Ing. Fernando N. Dulout" (UNLP-CONICET LA PLATA), Facultad de Ciencias Veterinarias UNLP, La Plata, Argentina, **2** Departamento de Genética, Universidad de Córdoba, Córdoba, España, **3** Department of Animal Science, University of California, Davis, Davis, California, United States of America, **4** Departamento de Producción Animal, Facultad de Ciencias Veterinarias, Universidad Nacional de La Plata, La Plata, Buenos Aires, Argentina

* sdemyda@igevet.gob.ar

OPEN ACCESS

Citation: Goszczynski D, Molina A, Terán E, Morales-Durand H, Ross P, Cheng H, et al. (2018) Runs of homozygosity in a selected cattle population with extremely inbred bulls: Descriptive and functional analyses revealed highly variable patterns. PLoS ONE 13(7): e0200069. https://doi.org/10.1371/journal.pone.0200069

Editor: Arda Yildirim, Gaziosmanpasa University, TURKEY

Received: September 18, 2017

Accepted: June 19, 2018

Published: July 9, 2018

Abstract

The analysis of runs of homozygosity (ROH), using high throughput genomic data, has become a valuable and frequently used methodology to characterize the genomic and inbreeding variation of livestock and wildlife animal populations. However, this methodology has been scarcely used in highly inbred domestic animals. Here, we analyzed and characterized the occurrence of ROH fragments in highly inbred (HI; average pedigree-based inbreeding coefficient F_{PED} = 0.164; 0.103 to 0.306) and outbred Retinta bulls (LI; average F_{PED} = 0.008; 0 to 0.025). We studied the length of the fragments, their abundance, and genome distribution using high-density microarray data. The number of ROH was significantly higher in the HI group, especially for long fragments (>8Mb). In the LI group, the number of ROH continuously decreased with fragment length. Genome-wide distribution of ROH was highly variable between samples. Some chromosomes presented a larger number of fragments (BTA1, BTA19, BTA29), others had longer fragments (BTA4, BTA12,

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 18 2013, pages 2355–2356
doi:10.1093/bioinformatics/btt398

Sequence analysis

Advance Access publication July 9, 2013

BioSmalltalk: a pure object system and library for bioinformatics

Hernán F. Morales* and Guillermo Giovambattista

Instituto de Genética Veterinaria (IGEVET), CONICET La Plata–Facultad de Ciencias Veterinarias, Universidad Nacional de La Plata, La Plata B1900AVW, CC 296 Argentina

Associate Editor: Janet Kelso

ABSTRACT

Summary: We have developed BioSmalltalk, a new environment system for pure object-oriented bioinformatics programming. Adaptive end-user programming systems tend to become more important for discovering biological knowledge, as is demonstrated by the emergence of open-source programming toolkits for bioinformatics in the past years. Our software is intended to bridge the gap between bioscientists and rapid software prototyping while preserving the possibility of scaling to whole-system biology applications. BioSmalltalk performs better in terms of execution time and memory usage than Biopython and BioPerl for some classical situations.

Availability: BioSmalltalk is cross-platform and freely available (MIT license) through the Google Project Hosting at http://code.google.com/p/biosmalltalk

Contact: hernan.morales@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 12, 2013; revised on June 5, 2013; accepted on July 3, 2013

1 INTRODUCTION

We present a novel free/open source software (FOSS) platform for the development of bioinformatics software and applications. BioSmalltalk attempts to reconcile the current *de facto* scripting modalities of textual programming languages with the features of Smalltalk (Goldberg and Robson, 1983), which has a pure object dynamic programming environment.

BioSmalltalk provides similar functionality to other FOSS toolkits for bioinformatics, such as BioPerl (Stajich *et al.*, 2002), Biopython (Cock *et al.*, 2009) and BioJava (Holland *et al.*, 2008), based in industry-leading general-purpose textual programming languages, and toolkits, including the Bio* projects. The Bio* toolkits' usage of OO is commonly hybrid or emulated through modules (Cock *et al.*, 2009; Stajich *et al.*, 2002), mixing objects with primitive data types and hampering the use of reflective functionalities (Maes, 1977). BioSmalltalk benefits from decreased source code verbosity, and its execution in a self-contained snapshot system that promotes run-time adaptability, critical for systems where shutdown cycles cannot be tolerated (Hirschfeld and Lämmel, 2005).

2 FEATURES

2.1 Bioinformatics

BioSmalltalk provides objects to manipulate biological sequences and data from databases like the Entrez system (Schuler *et al.*, 1996). It also contains wrappers for command-line tools like ClustalW (Thompson, 1994) and HMMER (Finn, 2011) sequence visualization and format conversion.

We based implementation on existing FOSS bioinformatics platforms, specifically BioPerl and Biopython, to prevent educational obsolescence, preserving the familiar object model interfaces for experienced bioinformaticians.

BioSmalltalk contains tokenizers, parsers and formatters for common sequence identifiers, FASTA, BLAST and Entrez XML, PHYLIP (Felsenstein, 1989), Arlequin (Excoffier, 2005) and others. Most parsers use PetitParser (Renggli *et al.*, 2010), a dynamically reconfigurable parser library. Additional features can be found in the project documentation. We did a microbenchmark to compare the performance of our library using the script in Figure 1. We have executed the scripts five times immediately after booting without unnecessary processes (Tests were performed on GNU/Linux Debian kernel 2.6.32-

https://github.com/hernanmd/BioSmalltalk

Thank you