Inter-American Development Bank

Bioinformatics for Researchers in Life Sciences: Tools and Learning Resources

Institutions for Development Sector

Competitiveness, Technology, and Innovation Division

DATABASE/DATASET

IDB-DT-66

Author: Natalia S. Rodriguez Muxica
Editor: Rafael Anta

February 2022

Bioinformatics for Researchers in Life Sciences: Tools and Learning Resources

Author: Natalia S. Rodriguez Muxica
Editor: Rafael Anta

Inter-American Development Bank
Competitiveness, Technology, and Innovation Division

February 2022

IDB-DT-66

“Bioinformatics for Researchers in Life Sciences: Tools and Learning Resources”

Methodology Report

November 2021

Natalia S. Rodriguez Muxica
Rafael Anta

INTRODUCTION

The experience of the COVID-19 pandemic has shown that bioinformatics, the intersection of biology and computer science, has a fundamental role in scientific research strategies in all the disciplines involved in fighting COVID-19: epidemiology, virology, biomedicine, genomics and synthetic biology, among others. This experience also suggests that bioinformatics skills in Latin America are incipient and that its scientific systems cannot take full advantage of the tools and data available thanks to the rise of open science and collaboration through global networks. In this context, CTI commissioned a study to produce a catalog of bioinformatics tools and related training that may help stimulate and encourage the development of bioinformatics skills in the region.

OBJECTIVES

General Objective

Investigate which are the main digital skills required in bioinformatics to work in the response to the COVID-19 pandemic, and analyze the availability of resources for the development of these skills.

Specific Objectives

1. Identify the main software and database tools used in the scientific disciplines of epidemiology, virology, biomedicine, and synthetic biology; and its associated costs.
2. Identify on the inventory of software tools and databases which digital skills are basic for any scientist and which digital skills are advanced.
3. Identify the availability of resources to learn how to use these tools and databases, and associated costs.
4. Identify skills necessary for the development of new bioinformatics applications, and the adoption of bioinformatics applications developed by others.

METHODOLOGY

The project "*Bioinformatics skills to face the COVID-19 Pandemic: Tools and Learning Resources*" was executed between June 2020 and October 2021 in four main phases:

1. Search and identification of relevant software for researchers in life sciences
2. Search and identification of relevant learning resources in bioinformatics
3. Review and enrichment of both catalogs based on feedback
4. Identification of skills necessary for the development of new bioinformatics applications, and for the adoption of bioinformatics applications by researchers from Latin America.

Search and identification of relevant software for researchers in life sciences

The tools gathered in this catalog are derived from three main sources: previously known tools by the author of the study, recommendations by researchers from Latin America, online inquiries both in spanish and in english, and based on the tools recommended or teached as part of the curriculum of the learning resources gathered in the second catalog.

Based on an initial exploratory search for tools and similar catalogs, the methodology and catalog template were established, defining the information needed from each tool, the categories used to classify them, and the selection criteria. All tools included in the final catalog were required to be available through the internet at the time this project was finished (either as a web-based tool or downloadeable), to be accessible either with or without any charge, and finally, they were not required to be actively supported or updated as long as they continued to be functional.

The following are the variables used in the catalog of bioinformatic tools:

- Tool_Name: Brand name or frequently used name for the tool.
- ID: Identification number for reference and internal use.
- Main_URL: Uniform resource locator (web address) to the tool or platform or to its provider.
- Usage: Main function of the tool or intended usage of the platform.
    - Categories:
        - Data Analysis: tools dedicated to all kinds of data analysis or to support analytical functions.
        - Visualization: tools whose main function is to generate visual representations and graph data.
        - Repository and DB: tools and organizations dedicated to providing access to digital information and/or physical resources.
        - Data Storage: tools dedicated to facilitating the storage of research data.
        - Communication and Community: tools dedicated to foster communication within the scientific community.
        - Marketplace and Collaboration: tools and platforms that facilitate access to resources, services and for peer collaboration.
        - Lab and Resource Management: tools for internal management of inventories, resources, records and other laboratory management activities.
        - Other: for tools that do not fit into one of the previous categories, or that collect several tools at once.
- Biotech_Fields: Fields within the biotechnology and life science sector that make use of the tool or that benefit directly from these resources.
    - Categories:
        - All
        - Biomathematics/Bioinformatics/Computational Biology
        - Biostatistics
        - Genetic Engineering
        - Genomics
        - Transcriptomics
        - Proteomics
        - Metabolomics
        - Cell Biology
        - Molecular Biology
        - Microbiology
        - Biochemistry
        - Biophysics
        - Physiology
        - Neurobiology
        - Immunology
        - Pharmacology
        - Epidemiology
        - Oncology
        - Biomedicine

        - Zoology
        - Marine Biology
        - Ecology
        - Evolution
        - Environmental Biology
        - Chemistry
        - Systems Biology

- Description: Detailed description and complementary information about the tool or platform. *This information has been collected from various sources and edited as needed.

- Technology_Type: Type of technological platform used to provide the tool.
    - Categories:
        - Web based/Cloud App: for platforms accessible on-line.
        - Desktop/Mobile App: for platforms with a downloadable desktop or mobile application option.
        - Programming Environment/Library: for tools provided as downloadable scripts or open-access libraries.
        - Other: for non-identified tools.

- Level_of_Complexity: "Combined level of complexity of the tool or platform itself and of the required knowledge expected to effectively handle it. This is an arbitrary distinction made to distinguish tools that require general information and skills available for most life science undergrads (tagged as ""Basic""), from those tools that require either further programing and computational skills or specialized knowledge in a field (tagged as ""Advanced''). Tools with both ""Basic"" and ""Advanced"" levels of complexity have been tagged as ""Basic to Advanced"". The tools where the level of complexity was not clear were also included in the ""Basic to Advanced"" category."
    - Categories:
        - Basic
        - Advanced
        - Basic to Advanced

- Pricing_Model: Main pricing model type of the tool or platform.
    - Categories:
        - Free / Open Source
        - Freemium
        - Subscription only
        - Usage based
        - Base plus overage
        - One-time payment
        - Other pricing model

- Pricing_Detail: "Detailed information about the cost of accessing the tool or platform and possible conditions applied according to the user's sector (academia or industry) or other factors. *Values as of June 2021. "

- Documentation_and_Tutorials: Supplemental information about the tool functionalities, installation manuals, usage guides and other tutorials, for those cases where it was available.
- Alternative_URLs: Alternative sources for related scientific publications or secondary platforms that provide access to the tool.
- Notes: Supplemental notes and comments for potential users.
- Year_Published: Year the tool was published. Based on their registered date in scientific publications, official website information or alternative sources such as source code repositories.
- Country_or_Region: Country or region where the tool was developed, based on authors affiliations and/or a company's country of operations.

Search and identification of relevant learning resources in bioinformatics

The learning resources gathered in this catalog are derived from three main sources: previously known training opportunities by the author of the study, from tutorials and training recommended by the tool's official websites, general online inquiries both in spanish and in english, and direct consultation in online learning platforms or relevant communities: the International Society for Computational Biology (ISCB), iGEM, Coursera, EdX, Udemy, European Molecular Biology Laboratory (EMBL), Swiss Institute of Bioinformatics (SIB), and open courseware from universities, among others.

As well as for the catalog of tools, based on the initial exploratory search the methodology and catalog template were established, defining the information needed from each learning resource, the categories used to classify them, and the selection criteria. All resources included in the final catalog were required to be active or imparted between the period of 2020 and 2021, to be available either online, presentially or in a hybrid format, and to be accessible either with or without any charge.

As a criteria the learning and training opportunities considered were those that did not require significant technical background or a level of education higher than a bachelor degree, and that were shorter and simpler than specialized undergraduate, graduate, and postdoctoral studies, and, finally, those resources that were accessible for life science professionals without requiring full time dedication for long periods of time and that could be complimentary with their current job responsibilities.

The following are the variables used in the catalog of bioinformatics learning resources:

- Resource_Name: Name of the program.
- ID: Identification number for reference and internal use.

- Institution_and/or_Platform: Institution or platform that provides the particular program.
- Main_URL: Main URL to access the program itself or the information to register.
- Resource_Type: Type of learning resource according to the amount of instructive hours, depth of the content delivered and format (written, video, interactive, etc.).
    - Categories:
        - Diploma
        - MOOC/Specialization
        - Course
        - Workshop
        - Webinar/Video Tutorial
        - Manual/Written tutorial
        - Other
- Modality: Modality used to deliver the contents.
    - Categories:
        - Online
        - Presential
        - Hybrid
- Level: Depth of the content being delivered through the program.
    - Categories:
        - Basic usage information
        - Introductory to the field/topic
        - Advanced training
- Content_Details: Details about the content and curriculum of the program, as described by the institution or platform providing it. For resources in Spanish the original language has been maintained. (This information has been collected from various sources and edited as needed).
- Availability: The frequency format in which the program is delivered.
    - Categories:
        - Sporadic
        - Timed (yearly/monthly/etc)
        - On demand
- Date_or_Duration: Date when the program is being imparted, for those that are sporadic or timed with a certain frequency. Duration of the program, for those cases that are delivered on demand.
- Language: Languages available for the program, either spoken or with the support of subtitles.
    - Categories:
        - Español
        - Português
        - English

- Cost: Official cost of the program. *Provided in the local currencies, and as of June 2021.
- Costs_Details: Details about the cost breakdown, exceptions, preferential rates and other information when needed.
- Biotech_Fields: Fields within the biotechnology and life science sector that make use of the tool or that benefit directly from these resources.
    - Categories:
        - All
        - Biomathematics/Bioinformatics/Computational Biology
        - Biostatistics
        - Genetic Engineering
        - Genomics
        - Transcriptomics
        - Proteomics
        - Metabolomics
        - Cell Biology
        - Molecular Biology
        - Microbiology
        - Biochemistry
        - Biophysics
        - Physiology
        - Neurobiology
        - Immunology
        - Pharmacology
        - Epidemiology
        - Oncology
        - Biomedicine
        - Zoology
        - Marine Biology
        - Ecology
        - Evolution
        - Environmental Biology
        - Chemistry
        - Systems Biology
- Tools_Addressed: References to the tools in the "Repository of Digital Tools for Researchers in Life Sciences" that are used during the program or that are related to the contents imparted according to the curriculum.
- Addittional_URLs: Alternative sources for related information about the course or other courses provided by the same organization.

Review and enrichment of both catalogs based on feedback

This phase consisted in iterative reviews to incorporate, refine and update all the information required based on the formats defined previously for both catalogs.

Identification of skills necessary for the development of new bioinformatics applications, and for the adoption of bioinformatics applications by researchers from Latin America

This final phase consisted first in a preliminary analysis of the tools and learning resources gathered in the previous steps in order to identify valuable insights, challenges and opportunities regarding the availability of digital tools as well as the availability and accessibility of training opportunities for them to be used by life science professionals in Latin America. This analysis also included the identification of tools used in the fields of greatest interest for the study: Epidemiology, Virology, Biomedicine, Synthetic Biology (Genetic Engineering & Systems Biology); and the identification of basic and advanced bioinformatic skills. The aforementioned analysis used as selection criteria the intersection of the two variables “Biotech Fields of Relevance” and “Level of Complexity", considering as a basic skill the use of those tools classified both as "Biotech Fields of Relevance” equal to “All" and “Level of Complexity" equal to "Basic". This criterion allows us to consider tools that can be useful for all fields of life sciences and that in turn do not require advanced digital skills (such as programming languages, use of bash/unix, etc.) or advanced knowledge of a particular discipline.

Secondly, a focus group was conducted in order to further explore the skills necessary to the development of new bioinformatic applications by professionals from Latin America, as well as to understand the current trends and challenges in this field inside the region. For this purpose, a group of bioinformatic professionals from the Latin American RSG-ISCB network was invited to participate in a discussion session on the 7th of June 2021, where a total of 11 members participated, representing 7 countries from the region (Chile, Peru, Colombia, Argentina, Mexico, Venezuela, and Ecuador). All insights gathered from this instance were summarized in the final presentation titled “*Final Results IDB Project: Bioinformatic Skills - 2021 - Convergencia BID _EN.pdf*”.

RESULTS

The project “*Bioinformatics skills to face the COVID-19 Pandemic: Tools and Learning Resources*” resulted in two main products:

5. A Catalog of Software for Researchers in Life Sciences
6. A Catalog of Learning Resources in Bioinformatics for Researchers in Life Sciences

Both results of the study can be found as part of the final package released by the Inter-American Development Bank in December 2021, together with this report, under the titles

"*Bioinformatics-Tools-for-LatAm-1-Catalog-of-Digital-Tools-for-Researchers-in-Life-Sciences-v4.0*" and "*Bioinformatics-Tools-for-LatAm-2-Catalog-of-Learning-Resources-in-Bioinformatics-v2.0.csv*". A final presentation of all results and insights derived from the study can be found in the same release package under the title "*Final Results IDB Project_ Bioinformatic Skills - 2021 - Convergencia BID _EN.pdf*".

ACKNOWLEDGEMENTS

The current project was commissioned by the Inter-American Development Bank, under the department of Institutions for Development and its Competitiveness Technology and Innovation Division. All content was created between June 2020 – October 2021, and published in December 2021. Author: Natalia S. Rodriguez Muxica; editor: Rafael Anta.

Metadata Sheet

Field	Description
Dataset Title	[EN] Bioinformatics Tools: A Catalog of Software for Researchers in Life Sciences. [ES] Herramientas Bioinformáticas: Un catálogo de software para investigadores en ciencias de la vida.
Description	The experience of the COVID-19 pandemic has shown that bioinformatics, the intersection of biology and computer science, has a fundamental role in scientific research strategies in all the disciplines involved in fighting COVID-19: epidemiology, virology, biomedicine, genomics and synthetic biology, among others. This experience also suggests that bioinformatics skills in Latin America are incipient and that its scientific systems cannot take full advantage of the tools and data available thanks to the rise of open science and collaboration through global networks. In this context, CTI commissioned a study to produce a catalog of bioinformatics tools and related training that may help stimulate and encourage the development of bioinformatics skills in the region. The dataset is a catalog of bioinformatics software for researchers and professionals working in life sciences. It includes tools for varied uses such as: data analysis, visualization, repositories and databases, data storage services, scientific communication, marketplace and collaboration, and lab resource management. Most tools are available as Web-based or desktop applications, while others are programming libraries. This catalogue includes more than 300 different tools, together with 10+ suggested entries for other third-party repositories that could be of use.
Dataset Type	Tables in CSV files
Publisher/Editor	Rafael Anta Competitiveness, Technology and Innovation Division (CTI) Institutions for Development (IFD)
Creator Contact Email	Natalia S. Rodriguez Muxica, research consultant, norodrig@uc.cl Rafael Anta, rafaela@iadb.org
Creation Date	June 2020 – October 2021
Time Period Covered	2020 – 2021
Country (s)	Global

Background Document Attachments	N/A
Variables Names and brief description	− Tool_Name: Brand name or frequently used name for the tool. − ID: Identification number for reference and internal use. − Main_URL: Uniform resource locator (web address) to the tool or platform or to its provider. − Usage: Main function of the tool or intended usage of the platform. ○ Categories: ● Data Analysis: tools dedicated to all kinds of data analysis or to support analytical functions. ● Visualization: tools whose main function is to generate visual representations and graph data. ● Repository and DB: tools and organizations dedicated to providing access to digital information and/or physical resources. ● Data Storage: tools dedicated to facilitating the storage of research data. ● Communication and Community: tools dedicated to foster communication within the scientific community. ● Marketplace and Collaboration: tools and platforms that facilitate access to resources, services and for peer collaboration. ● Lab and Resource Management: tools for internal management of inventories, resources, records and other laboratory management activities. ● Other: for tools that do not fit into one of the previous categories, or that collect several tools at once. − Biotech_Fields: Fields within the biotechnology and life science sector that make use of the tool or that benefit directly from these resources. ○ Categories: ● All ● Biomathematics/ Bioinformatics/ Computational Biology ● Biostatistics ● Genetic Engineering ● Genomics ● Transcriptomics ● Proteomics ● Metabolomics ● Cell Biology ● Molecular Biology ● Microbiology ● Biochemistry ● Biophysics ● Physiology ● Neurobiology

	● Immunology ● Pharmacology ● Epidemiology ● Oncology ● Biomedicine ● Zoology ● Marine Biology ● Ecology ● Evolution ● Environmental Biology ● Chemistry ● Systems Biology − Description: Detailed description and complementary information about the tool or platform. *This information has been collected from various sources and edited as needed. − Technology_Type: Type of technological platform used to provide the tool. ○ Categories: ● Web based/Cloud App: for platforms accessible on-line. ● Desktop/Mobile App: for platforms with a downloadable desktop or mobile application option. ● Programming Environment/Library: for tools provided as downloadable scripts or open-access libraries. ● Other: for non-identified tools. − Level_of_Complexity: "Combined level of complexity of the tool or platform itself and of the required knowledge expected to effectively handle it. This is an arbitrary distinction made to distinguish tools that require general information and skills available for most life science undergrads (tagged as ""Basic""), from those tools that require either further programing and computational skills or specialized knowledge in a field (tagged as ""Advanced''). Tools with ""Basic"" and ""Advanced"" levels of complexity have been tagged as ""Basic to Advanced"". *The tools where the level of complexity was not clear were also included in the ""Basic to Advanced"" category." ○ Categories: ● Basic ● Advanced ● Basic to Advanced − Pricing_Model: Main pricing model type of the tool or platform. ○ Categories: ● Free / Open Source ● Freemium ● Subscription only ● Usage based ● Base plus overage ● One-time payment

	● Other pricing model − Pricing_Detail: "Detailed information about the cost of accessing the tool or platform and possible conditions applied according to the user's sector (academia or industry) or other factors. *Values as of June 2021. " − Documentation_and_Tutorials: Supplemental information about the tool functionalities, installation manuals, usage guides and other tutorials, for those cases where it was available. − Alternative_URLs: Alternative sources for related scientific publications or secondary platforms that provide access to the tool. − Notes: Supplemental notes and comments for potential users. − Year_Published: Year the tool was published. Based on their registered date in scientific publications, official website information or alternative sources such as source code repositories. − Country_or_Region: Country or region where the tool was developed, based on authors affiliations and/or a company's country of operations.
Number of variables	15
Copyright and CC Licenses	The results offered in this dataset are those compiled by the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.
Additional Notes	N/A

Metadata Sheet

Field	Description
Dataset Title	[EN] Bioinformatics Tools: A Catalog of Learning Resources for Researchers in Life Sciences. [ES] Herramientas Bioinformáticas: Un catálogo de recursos de aprendizaje para investigadores en ciencias de la vida.
Description	The experience of the COVID-19 pandemic has shown that bioinformatics, the intersection of biology and computer science, has a fundamental role in scientific research strategies in all the disciplines involved in fighting COVID-19: epidemiology, virology, biomedicine, genomics and synthetic biology, among others. This experience also suggests that bioinformatics skills in Latin America are incipient and that its scientific systems cannot take full advantage of the tools and data available thanks to the rise of open science and collaboration through global networks. In this context, CTI commissioned a study to produce a catalog of bioinformatics tools and related training that may help stimulate and encourage the development of bioinformatics skills in the region. The catalog of learning resources in Bioinformatics is a dataset of available training opportunities to acquire the knowledge needed to make use of bioinformatic and digital tools for professionals working in the diverse fields of life science. These resources include training in diverse formats, ranging from diplomas and specializations, to courses and tutorials. The resources are available in three main languages: English, Spanish and Portuguese, and both in on-line and in-person modalities. This catalogue includes 144 different trainings, together with 5 suggested entries for other third-party repositories with training opportunities that could be of use.
Dataset Type	Tables in CSV files
Publisher/Editor	Rafael Anta Competitiveness, Technology and Innovation Division (CTI) Institutions for Development (IFD)
Creator Contact Email	Natalia S. Rodriguez Muxica, research consultant, norodrig@uc.cl Rafael Anta, rafaela@iadb.org
Creation Date	June 2020 – October 2021
Time Period Covered	2020 – 2021

Country (s)	Global
Background Document Attachments	N/A
Variables Names and brief description	− Resource_Name: Name of the program. − ID: Identification number for reference and internal use. − Institution_and/or_Platform: Institution or platform that provides the particular program. − Main_URL: Main URL to access the program itself or the information to register. − Resource_Type: Type of learning resource according to the amount of instructive hours, depth of the content delivered and format (written, video, interactive, etc.). o Categories: ● Diploma ● MOOC/Specialization ● Course ● Workshop ● Webinar/Video Tutorial ● Manual/Written tutorial ● Other − Modality: Modality used to deliver the contents. o Categories: ● Online ● Presential ● Hybrid − Level: Depth of the content being delivered through the program. o Categories: ● Basic usage information ● Introductory to the field/topic ● Advanced training − Content_Details: Details about the content and curriculum of the program, as described by the institution or platform providing it. For resources in Spanish the original language has been maintained. (This information has been collected from various sources and edited as needed). − Availability: The frequency format in which the program is delivered. o Categories: ● Sporadic ● Timed (yearly/monthly/etc) ● On demand − Date_or_Duration: Date when the program is being imparted, for those that are sporadic or timed with a certain frequency. Duration of the program, for those cases that are delivered on demand.

	− Language: Languages available for the program, either spoken or with the support of subtitles. o Categories: ● Español ● Português ● English − Cost: Official cost of the program. *Provided in the local currencies, and as of June 2021. − Costs_Details: Details about the cost breakdown, exceptions, preferential rates and other information when needed. − Biotech_Fields: Fields within the biotechnology and life science sector that make use of the tool or that benefit directly from these resources. o Categories: ● All ● Biomathematics/ Bioinformatics/ Computational Biology ● Biostatistics ● Genetic Engineering ● Genomics ● Transcriptomics ● Proteomics ● Metabolomics ● Cell Biology ● Molecular Biology ● Microbiology ● Biochemistry ● Biophysics ● Physiology ● Neurobiology ● Immunology ● Pharmacology ● Epidemiology ● Oncology ● Biomedicine ● Zoology ● Marine Biology ● Ecology ● Evolution ● Environmental Biology ● Chemistry ● Systems Biology − Tools_Addressed: References to the tools in the "Repository of Digital Tools for Researchers in Life Sciences" that are used during the program or that are related to the contents imparted according to the curriculum. − Addittional_URLs: Alternative sources for related information about the course or other courses provided by the same organization.

Number of variables	16
Copyright and CC Licenses	The results offered in this dataset are those compiled by the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.
Additional Notes	N/A

Bioinformatics for Researchers in Life Sciences

Tools and Learning Resources

●●●

Final Results

IDB Convergence Platform

Natalia S. Rodriguez Muxica
IDB Consultant
June 8th 2021
**Adaptation: October-November 2021*

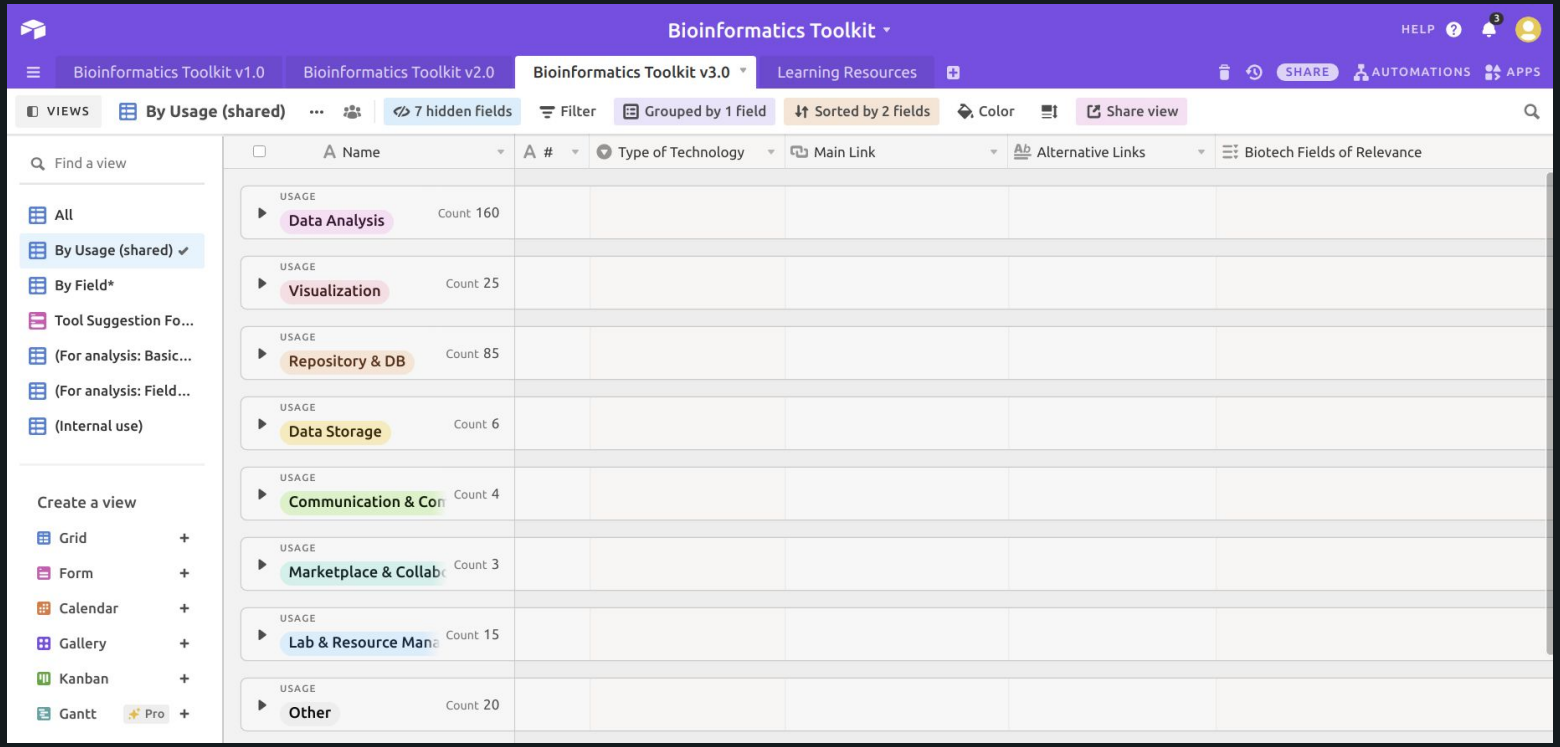
General Objective of the Study

Investigate which are the main digital skills required in bioinformatics to work in the response to the COVID-19 pandemic, and analyze the availability of resources for the development of these skills.

First specific objective

- Identify the main software and database tools used in the scientific disciplines of epidemiology, virology, biomedicine, and synthetic biology; and its associated costs.

Bioinformatics Tools Catalog



Tools Catalog: *general statistics, by type of usage*

302

Bioinformatic Tools

(+13 links to other repositories)

Past-1989:	**11** (4%)
1990-1999:	**28** (9%)
2000-2009:	**78** (25%)
2010-present:	**124** (39%)
N/A:	**74** (23%)

157
Data Analysis

85
Repository & DB

25
Visualization

15
Lab & Resource Management

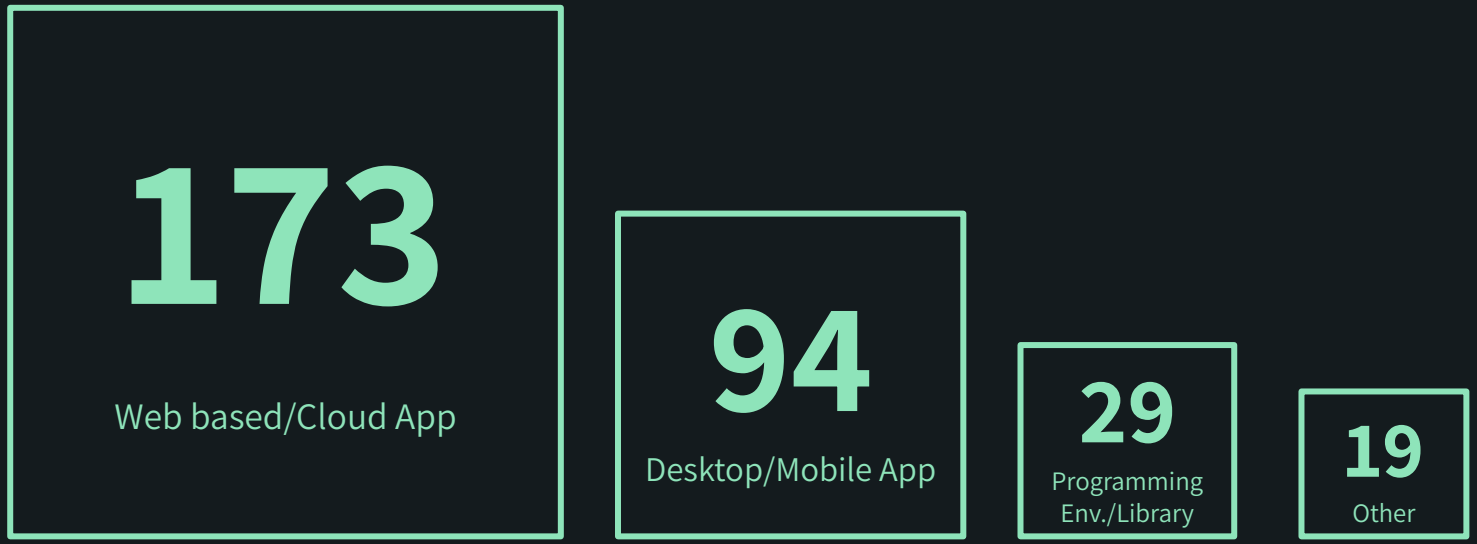
6
Data Storage

4
Communication & Community

3
Marketplace & Collaboration

20
Other

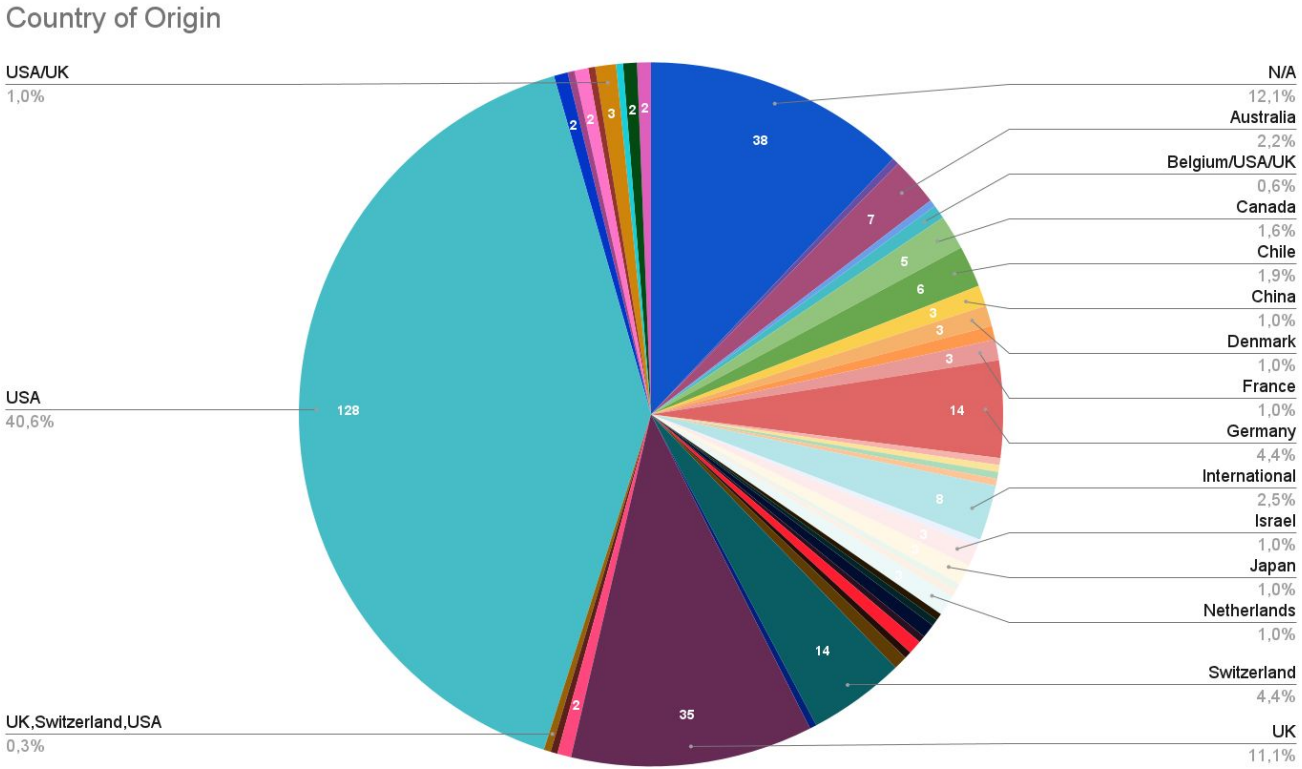
Tools Catalog: *general statistics, by type of technology*



Tools Catalog: *Fields of greatest interest for the study*

- **Tools by fields of interest:**
  - Epidemiology ⟶ 18
  - Virology = Microbiology ⟶ 65
  - Biomedicine ⟶ 21
  - Synthetic Biology = Genetic Engineering + Systems Biology ⟶ 187

Tools Catalog: *Countries of origin* *(* raw data)*



Tools Catalog: *Zoom-in by countries*

Country Sub-groups	% (#)
National Initiatives	**69,5%** (219)
US+	**48,3%** (152)
UK+	**16,2%**(51)
Switzerland+	**6,0%** (19)
Europe+	**34,6%** (109)
LATAM+	**2,5%** (8)
International Collaborations	**10,8%** (34)

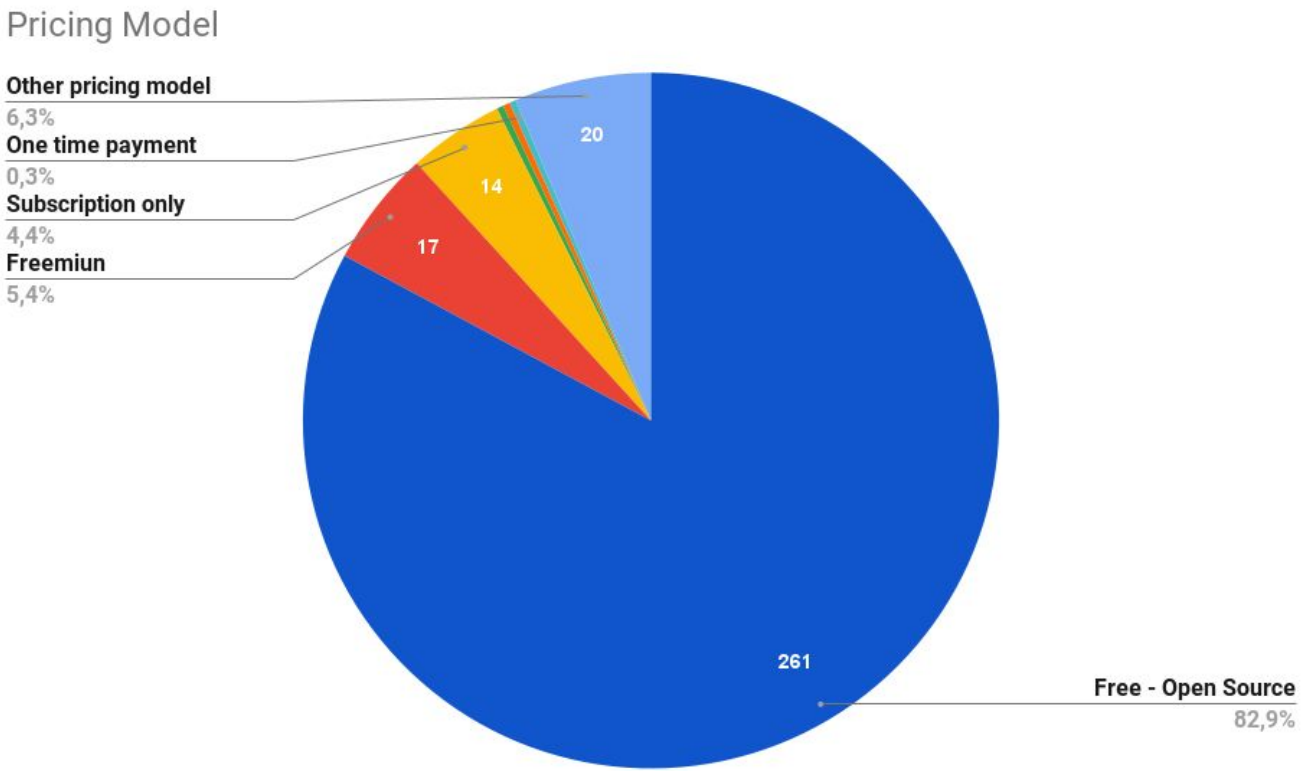
NCBI

BROAD INSTITUTE

EMBL-EBI

SIB Swiss Institute of Bioinformatics

Tools Catalog: *Associated costs*



Tools Catalog: *Associated costs*

278

With free options: Free + Freemium

17

Payed tools

Range in USD/year
$43-$50400

- Subscription only: 14
- Usage based: 1
- Base + overage: 1
- One time payment: 1

20

Variable or undetermined prices

Tools Catalog: *Associated costs, open-source challenges*

webact.org

This site can't be reached

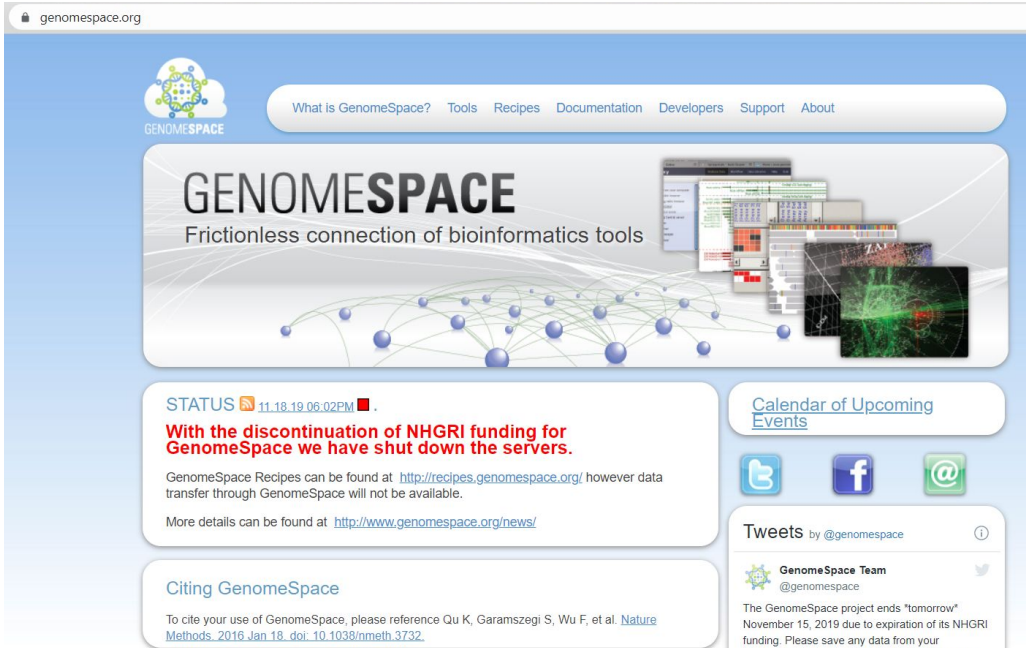
The connection was reset.

Try:

- Checking the connection
- Checking the proxy and the firewall

ERR_CONNECTION_RESET

Details



Second specific objective

- On the inventory of software tools and databases, identify which digital skills are basic for any scientist and which digital skills are advanced.

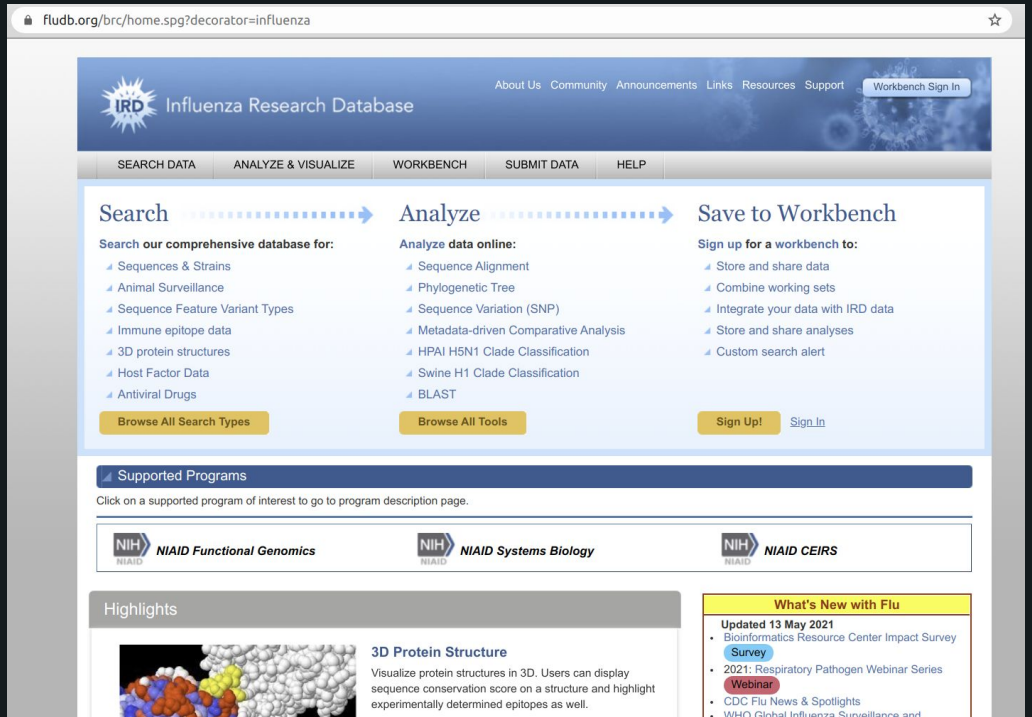
Tools: *Basic v/s advanced skills*

- **Skills that could be considered basic:**
  - → **73 tools meet this criterion**
  - Criterion used:
    - The tools considered as part of the “Basic Skills” were those in the set "Biotech Fields of Relevance”= “All" **+** “Level of Complexity" ="Basic"
    - This criterion allows us to consider those tools that can be useful for all fields of life sciences and that in turn do not require advanced digital skills (such as programming languages, use of bash / unix, etc.) or advanced knowledge of a particular discipline.
  - <u>Some general basic skills to highlight:</u>
    - Big data analysis with programming tools in R or Python
    - Use of user-friendly visualization and analysis tools: pymol / rasmol, Benchling / Snapgene
    - Direct and systematic consultation of databases
    - * IP analysis for particular technologies or sectors
    - * Internal management: digital lab notebooks, image repositories, and complementary data management tools

**non bioinformatic skills*

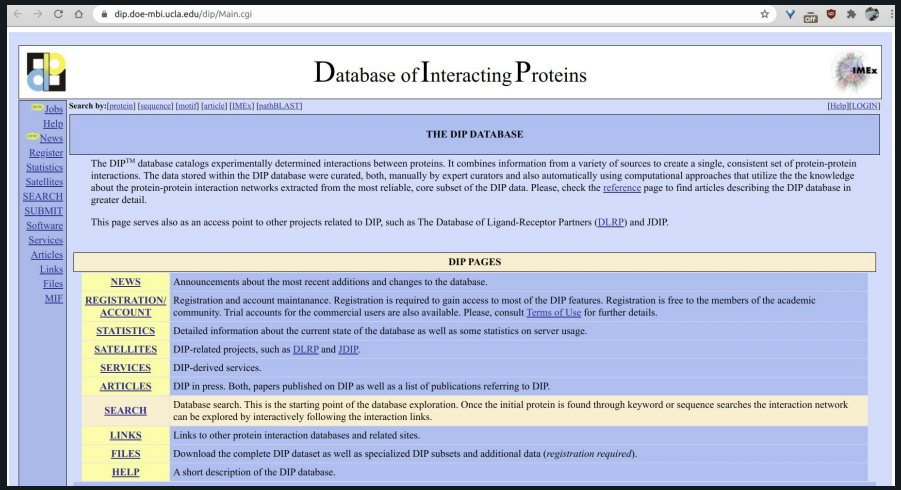
Tools: *Examples by field of interest*

- **Tools by fields of interest:**
  - Epidemiology → 18



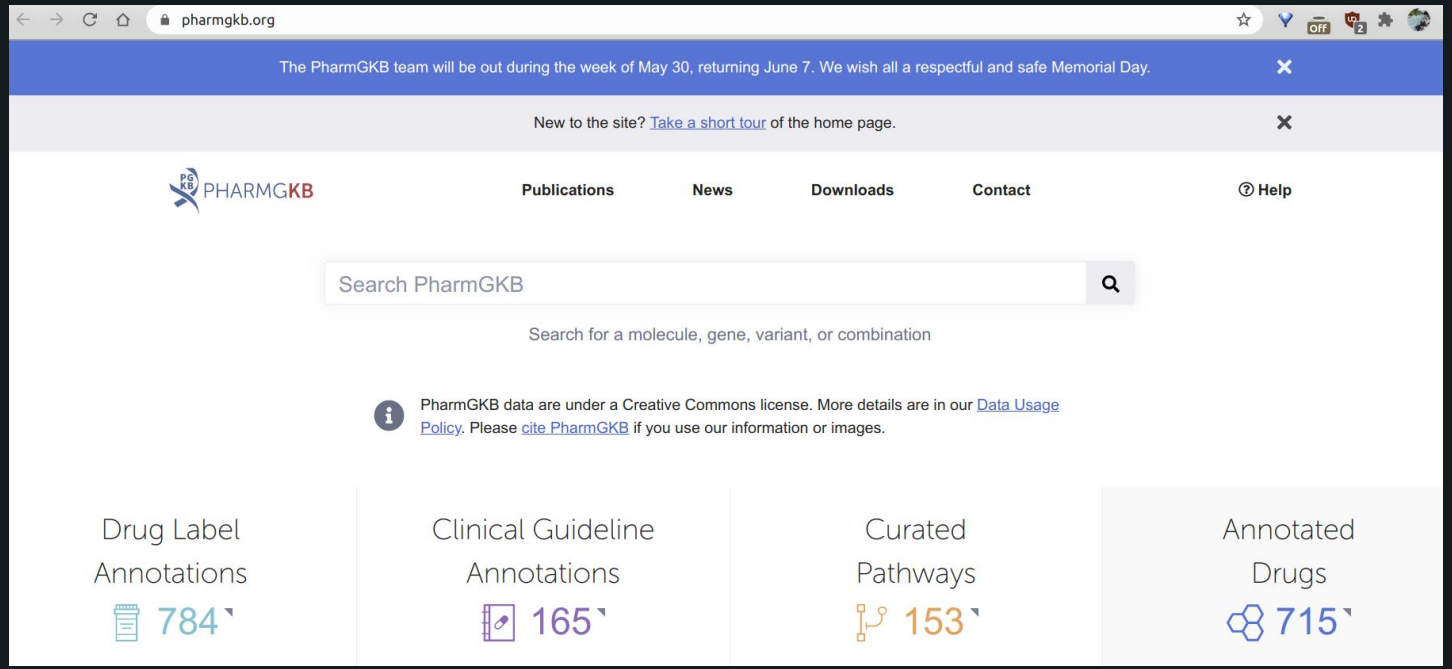
Tools: *Examples by field of interest*

- **Tools by fields of interest:**
  - Virology = Microbiology → 65



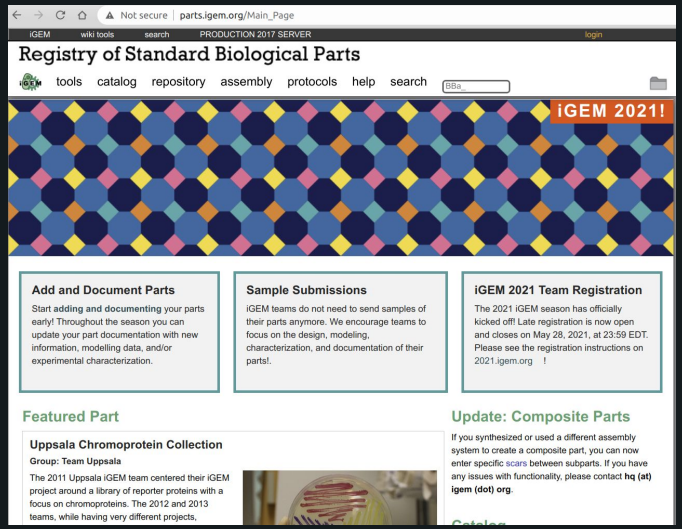
Tools: *Examples by field of interest*

- **Tools by fields of interest:**
  - Biomedicine → 21



Tools: *Examples by field of interest*

- **Tools by fields of interest:**
  - Synthetic Biology = Genetic Engineering + Systems Biology → 187



Tools Catalog: *Featured Insights*

- The universe of bioinformatics tools (which includes computational biology for these purposes) is constantly changing, integrating new tools of varying complexity and level of specialization, this in parallel with the development of its own sub-disciplines, as well as the growing digitization of scientific work in general.
- The tools contained within the catalog comprise mostly **data analysis tools (157/302)**, then general and specialized **database platforms (85/302)**, and thirdly **visualization tools (25/302)**. Then there are a couple of categories that include tools for the **management of internal tasks within laboratories and research teams (15/302)** that are of great importance for the traceability and internal recording of activities (such as Electronic Lab Notebooks, among others). And finally, some tools for **data storage** were included **(6)**, **communication tools for the community (4)** (such as ResearchGate), other tools for the purchase of products and **marketplace platforms for life sciences (3)**, and **others (20)**.
- The vast majority of **data analysis and visualization tools are part of a broader process or "pipeline"** within the investigation and study of a particular phenomenon, fulfilling very particular roles, interchangeable with other tools according to the particular needs of a study or the nature of the biological phenomenon studied.
- There are notable generational differences as of 2010, such as better usability for web tools, and open-source libraries available mostly through established repositories such as Github, Gitlab or Bitbucket.

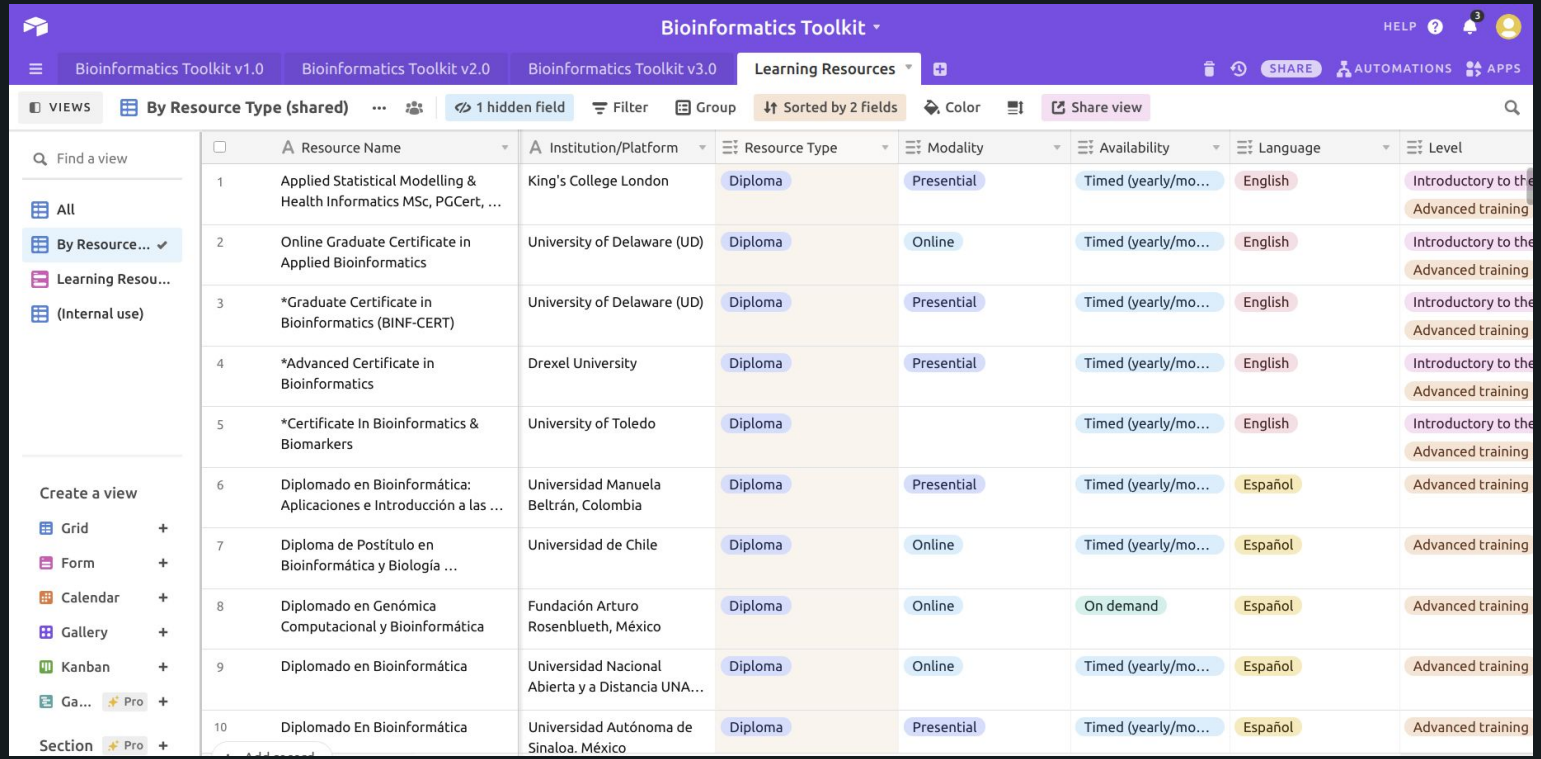
Tools Catalog: *Featured Insights*

- The countries involved in the generation of a greater part of these tools correspond in descending order to:
    - **The US involved in 48% of the 302 tools, the United Kingdom in 16%, and Switzerland in 6%.**
    - A large part of these tools are linked to institutions such as the NCBI, Broad Institute, EMBL-EBI and the Swiss Institute of Bioinformatics (SIB), which provide financing for their development and/or that host and maintain them within their clusters and digital platforms, helping to give continuity to this work.
- Regarding costs, there is a dominant **majority of tools that are free access or that allow free access for academic purposes: 278/302**.
- Trade-off between free access tools v/s paid tools:
    - Dedicated support and with the possibility of customization to the needs of use for paid tools.
    - **Less sustainability over time for open access tools:**
        - Updating of the tools is fragile, it is given by the availability and interest of its creators.
        - Web hosting is a limitation, some web tools are lost due to lack of funds to give them continuity. Successful cases are given by the availability of their own institutions or other external ones to give access to their servers (such as the EMBL, SIB, NCBI, universities or centers).

Third specific objective

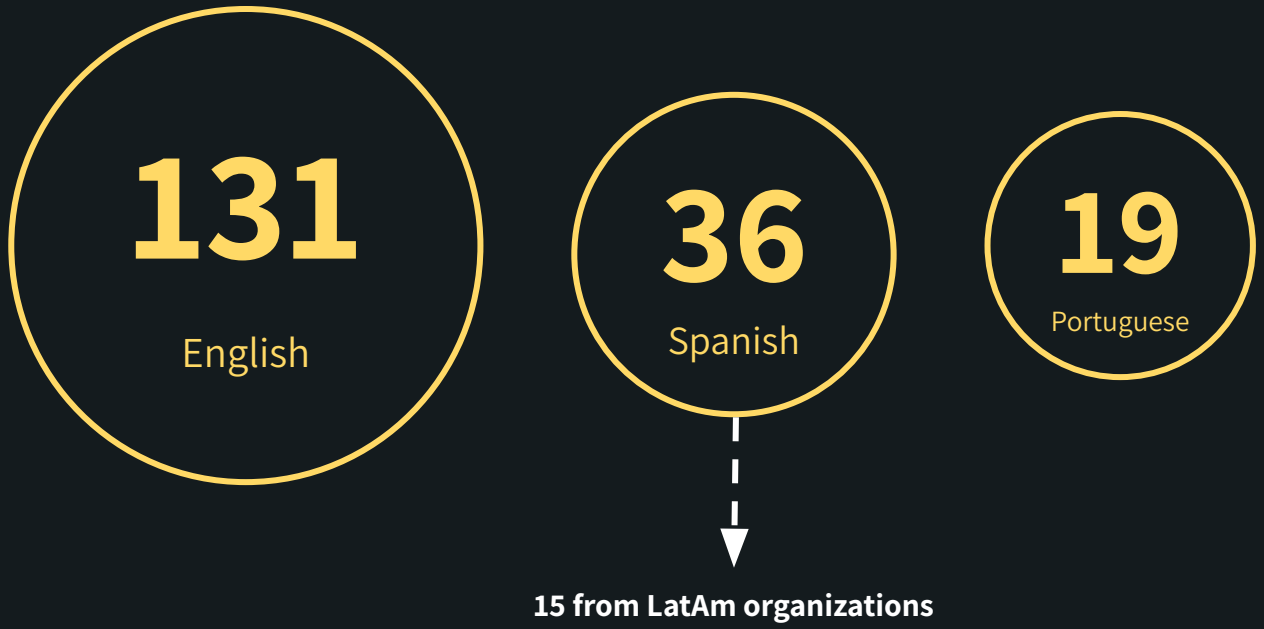
- Identify the availability of resources to learn how to use these tools and databases, and associated costs.

Learning Resources Catalog



Learning Resources Catalog: *Resource type*

Learning Resources Catalog: *Language*



Learning Resources Catalog: *Associated costs*

60/149

With free options: Free + Freemium

73

Payed resources

Range in USD
$12-$14.839

16

Undetermined prices

Learning Resources Catalog: *Featured Insights*

- **Most of the tools for data analysis and visualization coming from the first catalog are part of a process or "pipeline"** within the study of a particular phenomenon, and they fulfill very particular roles in this process, sometimes interchangeable with other tools according to to the particular needs of a study or to the nature of the biological phenomenon studied.
    - This reality motivated the need to generate a separate repository for learning resources since many of them must be taught together.
    - This catalog does not include the tutorials or specific user manuals that many of the tools have, they were included within the tools repository in the category "Documentation & Tutorials", with some exceptions for widely used tools such as R and Python.
- Through the availability of learning resources, **is possible to observe consolidated topics in the field that have extensive learning resources, and emerging topics that have a smaller offer of learning resources**:
    - Consolidated topics: statistics, genomics and analysis of sequencing data in general.
    - Emerging topics: metagenomics, systems biology, molecular simulations, search for new therapeutic targets or drugs, use of artificial intelligence applied to biomedicine.

Insights about Latin America (LA): *From both catalogs*

- In general, the region is underrepresented both in the development of bioinformatics tools and in the generation of quality learning resources focused on the needs of local scientists.

- Regarding bioinformatics tools:
  - **Authors from LA are involved in the development of only 2.5% of the tools in the catalog, with some cases of international collaborations and no relevant case of intra-regional collaborations.**
  - Most of the catalog tools developed by the region are not widely used by the global community, and they resolve minor technology gaps or for specific use cases.

- Regarding the learning resources available:
  - **A 24% (36/149) of the resources have options in Spanish, of which only 15 resources (including diplomas and courses) have been generated from LA itself.**
  - There are new and interesting cases such as iGEM teams that in order to raise funds are generating courses for the community, and organizations such as Bioscience App and Allbiotech that are also facilitating quality courses in spanish for the LA region.
  - There is a low availability of diploma programs, and local diploma programs address broad topics in a short time (1 year in part-time mode), with a majority focused on the analysis of genomic data, a smaller group on structural biology, and one particular case that includes systems biology.
  - The graduate diplomas are concentrated in Mexico, Colombia and Chile.

Fourth specific objective

- **Identify skills necessary for the development of new bioinformatics applications, and the adoption of bioinformatics applications developed by others.**
    - Focus group with bioinformatic professionals from the RSG-ISCB network of LatAm: June 7, 2021

*https://www.google.com/url?sa=i&url=https%3A%2F%2Fhesonline.com%2F2019%2F09%2F09%2Ffocus-groups-focusing%2F&psig=AOvVaw2OT7ngrOWTOSPZatzU7m9r&ust=1623174092817000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKDpuvCIhvECFQAAAAAdAAAAABAD

Insights from the Focus Group: *LatAm RSG-ISCB Network* *

- **Total number of participants: 11**
- **Countries represented: Chile, Peru, Colombia, Argentina, Mexico, Venezuela, and Ecuador.**

- **Regarding the digital and computer skills of life science scientists in general (non-bioinformatics), and the adoption of bioinformatics tools:**
  - Experimental scientists generally show a deficiency in the use of digital tools. This lack of digital skills includes those like handling Linux systems, console commands, or programming in languages like R or Python.
  - There is a generational gap and also a language barrier with English, which is the dominant language, to be able to use these tools and access information such as courses and tutorials.
  - There is mistrust in the potential and credibility of studies carried out purely *in silico*, which makes their adoption difficult.
  - Undergraduate bioinformatics courses fail to deliver basic skills such as programming, in general they only allow a first approach to theoretical or introductory uses of bioinformatics tools.
  - Proposal: Retrain or re-educate principal investigators who lead the teams and who are the ones who make a large part of the decisions regarding personnel hiring and the research methodologies to be used under their tutelage.

Insights from the Focus Group: *LatAm RSG-ISCB Network* *

- **Regarding the job market conditions for LatAm bioinformatic professionals and their training opportunities in the region:**
    - There is a lack of placings for bioinformaticians and research projects that make use of these tools. This is shared by the representatives of Peru, Colombia, Venezuela, Ecuador and Argentina, with the exception of Chile where it is perceived that there is a demand for bioinformatics professionals.
    - Systems for awarding scholarships, allocating funds, thesis project requirements, and others sometimes limit the use of bioinformatics as the primary method and / or the hiring of these professionals.
    - There has been a growth in the number of professionals who are trained in bioinformatics, and of new training programs within the region, however in many countries there are not enough job opportunities for all those who are trained.
    - Brazil, Mexico and Chile are the countries considered leaders in the region in research on the subject.
    - Argentina and Chile in particular, are pioneers in their training programs dedicated 100% to bioinformatics.
    - The lack of renewal of positions limits the entry of young professionals trained in these disciplines and of professionals with greater knowledge of their potential.
    - Bioinformaticians insertion in the private sector is still limited.
    - Bioinformaticians hired in some cases make up for other lacks in infrastructure and access to technical service at universities and centers that do not correspond to their positions.

Insights from the Focus Group: *LatAm RSG-ISCB Network* *

- **Regarding the technical infrastructure necessary to make use of bioinformatics tools in the different fields and at different levels of complexity:**
    - The necessary infrastructure is complex and expensive to maintain, which is not always visible to decision makers. Some limiting factors include:
        - Infrastructure for the storage of large amounts of data, and their corresponding backups.
        - Personnel in charge of maintenance.
        - Large quantity and quality of processors required for demanding analysis and complex studies.
        - Security systems to prevent equipment from being damaged.
        - Stable internet connection: Internet access is limited in certain countries, particularly outside capital cities, limiting access to tools and the use of local clusters from other regions.
        - Access to financing from public funds for computers, hard drives, and other electronics.
    - Investments have been investments made in clusters for academic use in Peru, Colombia and Ecuador, however they have not yielded the expected results due to factors such as the ones stated above, and due to the lack of knowledge/awareness of the community and their own institutions on how to get the most out of these facilities.
    - The development of tools is still led by developed countries because it requires resources such as: high-power clusters and research funds that give researchers the freedom to create scientific results in the form of tools or information for the improvement of other analyzes, and not for necessarily for publication in scientific journals.
    - Proposal: Partner within the region to generate regional clusters that everyone can access through the internet.

Insights from the Focus Group: *LatAm RSG-ISCB Network **

- **Regarding the access to these tools and to training resources for the use of them:**
    - Very broad courses in bioinformatics in general tend to be superficial due to the broad scope of the field and they also fail to generate the necessary skills.
    - Available tutorials are mostly in English and cryptic as they do not explain the developed process in detail and assume the management of certain skills.
    - Proposals:
        - Teach programing as part of the base curriculum for scientific careers.
        - Include R in statistics course syllabi as a first step.
        - Create bootcamps in bioinformatics that in a period of only 6 months can deliver enough skills to professionals in the scientific area to be employable.

**All the "insights" shared here were part of the conversation carried out with bioinformatics professionals from Latin America and the Caribbean, belonging to the RSG-ISCB network. Thus, the information shared in this section reflects only the opinion of its authors based on their personal experiences in academia in their respective countries, and not a judgment exercised by those in charge of this study or by the IDB.*

Final comments:

***** The information and statistics provided in this presentation correspond to the study completed in June 2021.*

****** The information in both catalogs, made available through the Inter-American Development Bank platform, was updated between September and October 2021, including new tools based on user suggestions.*


IDB
Inter-American
Development Bank