Idoia Ochoa

Mikeletegi Pasealekua, 48, 20009, Donostia, Spain
☎ *+34 (616) 680 931*
✉ *idoia@illinois.edu, iochoal@unav.es*
🖱 ***http://idoia.ece.illinois.edu***

Last updated: March 2020

Position

University of Navarra - TECNUN *Spain*
Collaborator Professor ***01/2020–Present***

- Department of Electrical Engineering
- Institute of Artificial Intelligence (affiliate)

University of Illinois at Urbana-Champaign *IL, USA*
Assistant Professor ***01/2017–Present***

- Department of Electrical and Computer Engineering
- Department of Computer Science (by courtesy)
- Carl R. Woese Institute for Genomic Biology (affiliate)

Education

Stanford University *CA, USA*
PhD candidate in Electrical Engineering ***09/2010–09/2016***

- Thesis: Genomic data compression and processing: theory, models, algorithms and experiments.
  - Advisor: Prof. Tsachy Weissman (tsachy@stanford.edu)

Stanford University *CA, USA*
Master in Electrical Engineering, *GPA: 4.031* ***09/2010–12/2012***

TECNUN, University of Navarra *Spain*
Telecommunications Engineering Degree ***09/2003–02/2009***

- Ranking position: 2 out of 51.
- Master thesis: Iterative Decoding Techniques for the Relay Channel *(GPA: 10/10)*.
  - Advisor: Prof. Pedro M. Crespo (pcrespo@ceit.es)

Lulea Tekniska Universitet *Sweden*
Erasmus Program ***08/2007–01/2008***

Interests

Computational Genomics, Bioinformatics, Data Compression, Machine learning, Information Theory and Coding, Signal Processing.

Experience

Stanford University *CA, USA*
Group of Prof. Tsachy Weissman in the department of Electrical Engineering ***2010–2016***

I worked on the design and development of new algorithms to improve the distribution and storage of genomic data, to facilitate its access, and to boost the inferential power of analysis performed on it. My approach combines tools from information theory, statistics, and machine learning.

- Contributions:
  - Designed lossless and lossy compressors for genomic data
  - Developed a denoiser to reduce noise present in genomic data
  - Proposed a new method for cancer module discovery using gene expression data
  - Designed compression schemes for databases that enable efficient similarity queries in the compressed domain

Brown Hill Productions *CA, USA*
Technical consultant ***2014–2015***

- Supervisor: Producer Jonathan Dotan
- Technical consultant for the second season of HBO's TV show "Silicon Valley".

Genapsys *CA, USA*
Internship ***Summer 2014***

- Supervisor: Dariush Dabiri (dariush@genapsys.com)
- Design and implementation (in Matlab) of the base-calling algorithm for Genapsys' new sequencing technology.

Google *CA, USA*
Internship ***Summer 2013***

- Supervisor: Lynn Bos (lynnb@google.com)
- Design, implementation and verification of a flexible decimation filter for a digital microphone.
  - Architecture design in Matlab
  - Implementation of a C-model including fixed point conversion
  - C-to-RTL conversion using the tool Catapult
  - FPGA implementation

CEIT, Centre of Studies and Technical Research of Gipuzkoa *Spain*
Research Assistant in the Electrical Engineering Department ***2009–2010***

- Supervisor: Prof. Pedro M. Crespo (pcrespo@ceit.es)
- Conducted research in the relay channel, source-channel coding for non-uniform and cyclostationary sources, LDPC and Turbo codes.

Research Assistant in the Electrical Engineering Department ***2008***

- Supervisor: Prof. Andoni Irizar
- Design of a PLCP (Physical Layer Convergence Procedure) with VHDL, for a UWB (Ultra Wide Band) system.

TECNUN, University of Navarra *Spain*
Collaborator Student in the Electrical Engineering Department ***2005***

- Worked on Cadence circuit design and lay-out

Teaching experience

University of Navarra - TECNUN *Spain*
Collaborator Professor ***2020–Present***

- High Performance Computing *(Spring 2020)*
- Computer Architecture II *(Spring 2020)*

University of Illinois at Urbana-Champaign *IL, USA*
Assistant Professor ***2017–Present***

- Advanced Digital Communications *(ECE562, Fall 2018)*
- Probability with Engineering Applications *(ECE313 – MATH362, Spring 2017, Spring 2018, Spring 2019)*

Data Analysis master (titulo propio), University of Navarra - TECNUN *Spain*
Assistant Professor ***2019***

- Machine Learning *(Summer 2019)*

Girls' Adventures in Mathematics, Engineering, and Science camp, UIUC *IL, USA*
Assistant Professor ***2018***

- Signal Processing *(Summer 2018, Summer 2019)*

Stanford University *CA, USA*
Teaching Assistant ***2013–2015***

- Information Theory *(EE376A – STATS376A, Winter 2015, Winter 2016)*
- Universal Schemes in Information Theory *(EE376C, Spring 2014)*
- Introduction to Statistical Signal Processing *(EE278B, Winter 2014)*
- Inference, Estimation and Information Processing *(EE378A, Spring 2013)*

TECNUN, University of Navarra *Spain*
Teaching Assistant ***2003–2004***

- Laboratory of Circuits and Physics I

Research Grants

Awarded a **research grant from the ANII (Agencia Nacional de Investigacion e Innovacion)**, Uruguay (co-PI, UYU 1,000,000, 2 years) *2020*

Awarded a **research grant from the CSIC (Comision Sectorial de Investigacion Cientifica)**, Uruguay (co-PI, UYU 1,249,996, 2 years) *2019*

Awarded a **Strategic Research Initiative (SRI) Phase II grant** from the University of Illinois at Urbana-Champaign (UIUC) (main PI, $70,000, 1 year) *2019*

Awarded a **Strategic Research Initiative (SRI) Phase I grant** from the University of Illinois at Urbana-Champaign (UIUC) (main PI, $75,000, 1 year) *2018*

Awarded a **Chan Zuckerberg Initiative (CZI) grant**, under the Human Cell Atlas (main PI, $105,000, 1 year) *2018*

Awarded an **NIH grant**, under the BD2K initiative, in collaboration with the University *2015* of Illinois at Urbana-Champaign (UIUC)

Preprints and Under Review

- Q. Meng, **I. Ochoa**, and M. Hernaez, *GPress: a framework for querying General Feature Format (GFF) files and expression files in a compressed form*, Submitted to **Bioinformatics** (2nd review).

- C. Alberti, T. Paridaens, J. Voges, D. Naro, J. J. Ahmad, M. Ravasi, D. Renzi, P. Ribeca, G. Zoia, **I. Ochoa**, M. Mattavelli, J. Delgado, M. Hernaez, *An introduction to MPEG-G, the new ISO standard for genomic information representation*, 2018 (Bioarxiv).

- M. Goyal, K. Tatwawadi, S. Chandak, **I. Ochoa**, *DeepZip: Lossless Data Compression using Recurrent Neural Networks*, 2018 (Arxiv).

- J. Peng, **I. Ochoa**, and O. Milenkovic, E^2M*: A Deep Learning Framework for Associating Combinatorial Methylation Patterns with Gene Expression*, Submitted to **Bioinformatics** (Bioarxiv).

- A. No, M. Hernaez and **I. Ochoa**, *CROMqs: An Infinitesimal Successive Refinement Lossy Compressor for the Quality Scores*, Submitted to **IEEE/ACM Transactions on Computational Biology and Bioinformatics**.

Journal Papers

- I. Fisher-Hwang, **I. Ochoa**, T. Weissman, and M. Hernaez, *Denoising of Aligned Genomic Data*, **Nature Scientific Reports**, 2019.

- C. Zhang and **I. Ochoa**, *VEF: a Variant Filtering tool based on Ensemble methods*, **Bioinformatics**, 2019.

- J. Voges, T. Paridaens, F. Müntefering, L. S. Mainzer, B. Bliss, M. Yang, **I. Ochoa**, J. Fostier, J. Ostermann, and M. Hernaez, *GABAC: an arithmetic coding solution for genomic data*, **Bioinformatics**, 2019.

- R. Yang, X. Chen, and **I. Ochoa**, *MassComp, a Lossless Compressor for Mass Spectrometry Data*, **BMC Bioinformatics**, 2019.

- M. Hernaez, D. Pavlichin, and T. Weissman, **I. Ochoa**, *Genomic Data Compression*, **Annual Review of Biomedical Data Science**, 2019.

- S. Chandak, K. Tatwawadi, **I. Ochoa**, M. Hernaez, and T. Weissman, *SPRING: A next-generation compressor for FASTQ data*, **Bioinformatics**, 2018.

- L. Roguski, **I. Ochoa**, M. Hernaez, and S. Deorowicz, *FaStore – a space-saving solution for raw sequencing data*, **Bioinformatics**, bty205, 2018.

- J. Peng, O. Milenkovic, and **I. Ochoa**, *METHCOMP: A Special Purpose Compression Platform for DNA Methylation Data*, **Bioinformatics**, bty143, 2017.

- K. Tatwawadi, M. Hernaez, **I. Ochoa**, and T. Weissman, *GTRAC: fast retrieval from compressed collections of genomic variants*, **Bioinformatics**, btw437, 2016.

- **I. Ochoa**, M. Hernaez, R. Goldfeder, T. Weissman and E. Ashley, *Effect of lossy compression of quality scores on variant calling*, **Briefings in Bioinformatics**, bbw011, 2016.

- S. Deorowicz, S. Grabowski, **I. Ochoa**, M. Hernaez and T. Weissman, *Comment on: "ERGC: An efficient referential genome compression algorithm"*, **Bioinformatics**, btv704, 2015.

- G. Malysa, M. Hernaez, **I. Ochoa**, M. Rao, K. Ganesan and T. Weissman, *QVZ: lossy compression of quality values*, **Bioinformatics**, btv330, 2015.

- **I. Ochoa**, M. Hernaez and T. Weissman, *Aligned genomic data compression via improved modeling*, **Journal of bioinformatics and computational biology**, Vol. 12, No. 6, 2014.

- A. Manolakos, **I. Ochoa**, K. Venkat, A. Goldsmith and O. Gevaert, *CaMoDi: a new method for cancer module discovery*, **BMC genomics**, Vol. 15, No. Sup. 10, 2014 (**Highly Accessed**).

- **I. Ochoa**, M. Hernaez and T. Weissman, *iDoComp: a compression scheme for assembled genomes*, **Bioinformatics**, btu698, 2014.

- **I. Ochoa**, H. Asnani, D. Bharadia, M. Chowdhury, T. Weissman and G. Yona, *QualComp: a new Lossy Compression Algorithm of Quality Scores based on Rate Distortion Theory*, **BMC Bioinformatics**, Vol. 14, No. 1, 2013 (**Highly Accessed**).

- **I. Ochoa**, P. Crespo and M. Hernaez, *LDPC Codes for Non-Uniform Memoryless Sources and Unequal Energy Allocation*, **IEEE Communications Letters**, Vol. 14, No. 9, 2010.

- **I. Ochoa**, P. Crespo, J. Del Ser and M. Hernaez, *Turbo Joint Source-Channel Coding of Non-Uniform Memoryless Sources in the Bandwidth-Limited Regime*, **IEEE Communications Letters**, Vol. 14, No. 4, 2010.

Conference Papers

- G. Dufort y Alvarez, G. Seroussi, P. Smircich, J. Sotelo, **I. Ochoa**, and A. Martin, Compression of Nanopore FASTQ files, **7th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)**, Spain, 2019.

- **I. Ochoa**, H. Li, F. Baumgarte, C. Hergenrother, J. Voges and M. Hernaez, AliCo: a new efficient representation for SAM files, **Data Compression Conference (DCC)**, USA, 2019.

- R. Yang, X. Chen, and **I. Ochoa**, *MassComp, a Lossless Compressor for Mass Spectrometry Data*, **ISCB-LA SOIBIO EMBnet**, Chile, 2018.

- R. Long, M. Hernaez, **I. Ochoa** and T. Weissman, *GeneComp, a new reference-based compressor for SAM files*, **Data Compression Conference (DCC)**, USA, 2017.

- **I. Ochoa**, A. No, M. Hernaez, and T. Weissman, *CROMqs: an infinitesimal successive refinement lossy compressor for the quality scores*, **IEEE Information Theory Workshop (ITW)**, UK, 2016.

- K. Tatwawadi, M. Hernaez, **I. Ochoa**, and T. Weissman, *GTRAC: fast retrieval from compressed collections of genomic variants*, **European Conference on Computational Biology (ECCB)**, Netherlands, 2016.

- **I. Ochoa**, M. Hernaez, R. Goldfeder, T. Weissman and E. Ashley, *Denoising of Quality Scores for Boosted Inference and Reduced Storage*, **Data Compression Conference (DCC)**, USA, 2016.

- M. Hernaez, **I. Ochoa** and T. Weissman, *A cluster-based approach to compression of Quality Scores*, **Data Compression Conference (DCC)**, USA, 2016.

- **I. Ochoa**, M. Hernaez and T. Weissman, *Aligned genomic data compression via improved modeling*, **GIW ISCB-Asia**, Japan, 2014.

- A. Manolakos, **I. Ochoa**, K. Venkat, A. Goldsmith and O. Gevaert, *CaMoDi: a new method for cancer module discovery*, **GIW ISCB-Asia**, Japan, 2014.

- **I. Ochoa**, A. Ingber and T. Weissman, *Compression Schemes for Similarity Queries*, **Data Compression Conference (DCC)**, USA, 2014.

- **I. Ochoa**, A. Ingber and T. Weissman, *Efficient Similarity Queries via Lossy Compression*, **Allerton Conference on Communication, Control, and Computing**, USA, 2013.

- B. G. Chern, **I. Ochoa**, A. Manolakos, A. No, K. Venkat and T. Weissman, *Reference Based Genome Compression*, **IEEE Information Theory Workshop (ITW)**, Switzerland, 2012.

- **I. Ochoa**, P. Crespo, J. Del Ser and M. Hernaez, *Turbo Joint Source-Channel Coding of Cycle-Stationary Sources in the Bandwidth-Limited Regime*, The 2nd International Conference on Mobile Lightweight Wireless Systems (MOBILIGHT), Spain, 2010.

- M. Hernaez, P. Crespo, J. Del Ser and **I. Ochoa**, *Codigos LDGM concatenados para la transmision de fuentes correlacionadas para canales de difusion*, XXIV National Assembly of the International Union of Radio Science (URSI), Spain, 2009.

Conference Posters

- M. Goyal, K. Tatwawadi, S. Chandak, **I. Ochoa**, *DeepZip: Lossless Data Compression using Recurrent Neural Networks*, **Data Compression Conference (DCC)**, Utah, USA, 2019.

- C. Zhang, and **I. Ochoa**, *VEF: a Variant Filtering tool based on Ensemble methods*, **Intelligent Systems for Molecular Biology (ISCB-ISMB)**, Chicago, USA, 2018.

- R. Yang, and **I. Ochoa**, *MassComp, a Lossless Compressor for Mass Spectrometry Data*, **Rocky**

Mountain Bioinformatics Conference (ISCB), Colorado, USA, 2017.

- P. Li, M. Hernaez, **I. Ochoa**, and O. Milenkovic, *Micro: Microbial Database Compression Via Community Detection and Reference Discovery*, **Conference on Research in Computational Molecular Biology (RECOMB)**, San Diego, USA, 2016.

- **I. Ochoa**, M. Hernaez, R. Goldfeder, T. Weissman and E. Ashley, *Denoising of Quality Scores for Boosted Inference and Reduced Storage*, **Conference on Research in Computational Molecular Biology (RECOMB)**, San Diego, USA, 2016.

- K. Tatwawadi, M. Hernaez, **I. Ochoa**, and T. Weissman, *Fast retrieval from compressed collections of genomic variants*, **Conference on Research in Computational Molecular Biology (RECOMB)**, San Diego, USA, 2016.

Invited Talks

- **Moss, a novel multi-sample somatic variant caller**
  - CICbioGUNE, Derio, Spain, 2020.
- **MassComp, a Lossless Compressor for Mass Spectrometry Data**
  - ISMB/ECCB conference, Basel, Switzerland, 2019.
- **Genomic analysis pipeline: overview, challenges, and proposed solutions**
  - EUREKA, Science Museum, Donostia, Spain, 2020
  - MIT Math & CSAIL Bioinformatics Seminar, Massachusetts Institute of Technology (MIT), USA, 2019.
  - Universitat Politecnica de Catalunya, Spain, 2018.
- **MPEG-G Tutorial (Standard for Genomic Information Representation)**
  - L3/L3.1 meeting of the US National Body of the International Standardization Organization (ISO), Microsoft, USA, March 2018.
- **Denoising of Aligned Genomic Data**
  - The Seventh Spanish Workshop on Signal Processing, Information Theory and Communications, University of Navarra (TECNUN), Spain, January 2018.
- **Current challenges related to Genomic data processing and analysis**
  - Center for Science of Information (CSoI), Purdue University, Indiana, USA, April 2017.
- **Trends and Methods in Genomic Data Compression**
  - International Society for Computational Biology (ISMB), Orlando, USA, July 2016.

- **Genomic data compression, processing, and analysis**
  - University of Illinois at Urbana-Champaign, USA, April 2016.
  - Princeton University, USA, March 2016.
  - Washington University in St. Louis, USA, March 2016.
  - Stanford Compression Forum, USA, February 2016.

- **Genomic data compression**
  - Information Theory and Applications (ITA) workshop, San Diego, USA, February 2016.

- **Current trends on genomic data compression**
  - VA research seminar, Palo Alto, USA, November 2015.

- **Genomic Data Compression and Processing**

  - EECS Rising Stars, Massachusetts Institute of Technology (MIT), USA, November 2015.
- **An Overview of Genomic Data Compression**
  - Center for Science of Information (CSoI), Virtual Brown Bag Research Discussion Series, September 2015.
- **Compression schemes for similarity queries**
  - Stanford Compression Forum, Technical Talk, Stanford, USA, January 2015.
- **Efficient similarity queries via lossy compression**
  - New Directions in the Science of Information, Poster Presentation, UC Berkeley, USA, November 2014 **(second place winner)**.

Scholarships and Awards

Awarded the **MIT Innovators Under 35 Europe 2019**	*2019*
Selected to give a talk as a participant of the **MIT EECS Rising Starts Workshop**	*2015*
Richard and Naomi Horowitz Fellow, **Stanford Graduate Fellowship**	*2013–2015*
Basque Government Fellowship for graduate studies	*2013–2016*
Ranked 23^{rd} out of 147 candidates in the **EE Qualifying Examination** at Stanford	*2010*
La Caixa Fellowship Program to extend studies in the USA	*2010–2012*
Master Thesis funded by **Telefonica Fellowship**	*2008*

US Patents

- S.Chandak, K. Tatwawadi, T. Weissman, **I. Ochoa**, and M. Hernaez, *Systems and Methods for Compressing Genetic Sequencing Data*, US Patent Office, Application number 16545751, 02/20/2020

- A. Manolakos, **I. Ochoa**, K. Venkat, A. Goldsmith and O. Gevaert, *Cancer module discovery using gene expression data*, Submitted to Stanford's OTL - The Office of Technology Licensing, September 2014.

- **I. Ochoa** and M. Hernaez, *A Universal Compressor for Genomic Re-Sequencing Data*, Provisional US patent filled by Stanford's OTL - The Office of Technology Licensing, June 2014.

Service Activities

Workshop Organization Committees:

- Chair of special session on "Omics Data Processing and Analysis" at DSW (Data Science Workshop), Minneapolis, 2019.
- Chair of special session on "Omics Data Compression and Storage: Present and Future" at ISMB (International Society for Computational Biology), Chicago, 2018 (acceptance rate 20%).
- Chair of special session on "Bioinformatics" at the 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), October, 2018.
- Chair of special session on "Bioinformatics" at the 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), October, 2017.

Professional Organizations:

- Collins Scholar, University of Illinois at Urbana-Champaign (2017 - Present)

- International Organization for Standardization (ISO): Active participant in the initiative to define and establish a compression standard for genomic data (under the MPEG working group) (2013 - Present).
- Center for Science of Information (CSoI), NSF Science and Technology Center (2010 - Present)
- Stanford Compression Forum: organizer of the second edition (February 2016)
- Stanford Data Science Initiative (SDSI) (2014 - 2016)

Reviewer: Bioinformatics, Nature Technical Reports, Nature Biotechnology, BMC Bioinformatics, IEEE Communications Letters, several conference proceedings.

Additional Information

Languages: Native: Spanish, Proficiency: English, Low-Intermediate: German, French and Basque.

Computer skills: Programming Languages: C/C++, Python, Applications: R, MatLab, LATEX, MS Office, CVX, Java Operating Systems: Linux, UNIX, Windows.

Associations: President, Vicepresident and Financial Officer of Iberia - Spanish Association at Stanford (2010 - 2016).

References

Available upon request.