Hoan NGUYEN, PhD
Bioinformatics Analyst
The Jackson Laboratory
10 Discovery Driver, Farmington, CT 06032
Email : hoan.nguyen@jax.org/bmhoan@gmail.com
Phone: 860-994-3827
Web: https://www.jax.org

Research interests

- Bacterial genome assembly, genome comparative including SNP and methylation comparative
- Microbiome genomics and data integration
- Protein structure modeling, mutation discovery and prediction of the molecular consequences.
- NGS data analyses and translational bioinformatics design for genomics medicine.
- Big data infrastructure and workflow dedicated for computational science.

Education

2006 PhD in astrophysics (specialist: data-mining and virtual observatory), University of Strasbourg, France
2002 Master in computational science, University of La Rochelle, France
1997 B.S in mathematics and computer sciences, University of Hue, Vietnam

Experience

2015- Bioinformatics analyst, Jackson Laboratory, CT, USA
My research involves on further developing novel computational tools to investigate the structures of the microbiome and the role it plays in human health and disease. I provide the bioinformatics expertise and genome assembly for multiple lab's collaborators

2006-2014 Bioinformatics/Data Scientist, Institute of Genetics and Molecular and Cellular Biology (IGBMC), Strasbourg, France.
I led the development a comprehensive translational bioinformatics for the analysis of genetic variations involved in human diseases. My main research focused on evaluating the impact of identified mutation on the structure and function of a human protein. My work has been applied for discovering new mutations involved in human disease including neurological disorders, Age-related Macular and complete congenital stationary night blindness.

2002-2005 PhD Student, Centre National d'Etudes Spatiales (CNES) and Observatories Astronomiques de Strasbourg, Strasbourg, France
Development of SAADA System (Automatic Archival System for Astronomical Data allowing systematic exploitation of the European satellite XMM-Newton data catalogue, and of the International Virtual Observatory.

1999-2000 Training fellowship in education and computer science at Lille University, Lille, France

1998-1999 Teacher in computer science and applied mathematics at Hue University, Hue, Vietnam

Awards and Grants

2008-2012 The French Muscular Dystrophy Association (AFM) postdoctoral fellowship
2006-2007 The French National Center for Sientific Research postdoctoral fellowship
2002-2005 The French Government Space Agency (CNES) Ph.D. fellowship

Computational database and software contributions

MSV3d (variants database and protein stability change prediction): http://lbgi.fr/msv3d

SMP2H-CENTRAL (protein family analysis): http://lbgi.fr/sm2ph

KD4v (mutation prediction): http://lbgi.fr/kd4v

Gepetto-Gene Prioritization Toolkit : http://sourceforge.net/projects/gepetto/ (~700 downloads)

Saada: Astronomical database generator: https://sourceforge.net/projects/saada/ (~2000 downloads)

Oral Presentation at international conferences

MSV3d 2014: Database of human MisSense Variants mapped to 3D protein structure: Human Variome Project 5th Biennial meeting, May 2014, Paris, France.

Towards a Big Data Ecosystem for Translational Research. An application in human genetic variants, Oncotrans 2014. Reims,France.

Gepetto (GEne Prioritization ExTended Tool): An Open Source Framework for Gene Prioritization. 14th Annual Bioinformatics Open Source Conference BOSC 2013, Berlin, Germany July 19-20, 2013.

SM2PH-Central: An Integrative knowledgebase to investigate the genotype to phenotype relationships involved in human genetic diseases. Integrative Biology-2013. LA, USA, 2013**.**

Comprehensible Knowledge Discovery System for Missense Variant**,** Oral presentation at the 12th International Symposium on Mutation in the Genome,Lake Louise, Canada. April 2013.

Extracting Knowledge from a Mutation Database Related to Human Monogenic. Disease Using Inductive Logic Programming. International Conference on Bioinformatics, Computational Biology and Biomedical Engineering, Singapore, 2011.

How to publish local data into the Virtual Observatory with SAADA, Madrid, Spain, 2005.

Building an Astronomical database with SAADA, Madrid, Spain, 2005

SAADA: Astronomical databases made easier ADASS XIV, Caltech, CA, USA, 2004.

Reviewers for Journals:

Plos One

Briefings in Bioinformatics

BMC Medical genomics

IEEE/ACM Transactions on Computational Biology and Bioinformatics

Computers in Biology and Medicine Journal (Elsevier)

Languages:

Vietnamese (mother tongue)

French (fluently and confidently)

English

Bioinformatics skill

Genome assembly pipelines: Celera, PBcR, SMRTANALYSIS, SPADES,PBJelly.
Aligner: RTG, BWA, BOWTIE, BLASR,BLAST, LAST, NANOOK
Variant calling: SAMTOOLS, QUIVER
Classification tools: KRAKEN, RTG
NGS data processing: FastQC, SNAP,SAMtools, BWA,Bowtie, GATK, Tophat, Annovar, Galaxy
Copy Number Variant analysis: ExomeCNV, Pindel, Contra, VarScan, PennCNV
Transcriptomes and RNA-Seq analysis: R/ BioConductor packages
Structural characterization of mutant and homology search: I-Mutant, CSU, Modeler
Mutation prediction: Polyhen-2, SIFT, KD4v,VEP-Variant Effect Predictor
Protein family Analyses: Blast, DbClustal,Mafft, Macsims , TCofee.
Functional Annotation: GSEA, David, GO, HPO
Databases and repositories: Ensembl, UCSC/EnCode, Uniprot, Pfam, Genbank, PDB, dbSNP,ClinVar
Biological networks analysis: KEGG, Stringdb, Cytoscape JS
Libraries and API: NCBI API, Ensembl API, R, BioPerl, BioJava, BioPython.

Computer Science skill

Applied Statistical analyses and Machine Learning: Basic statistical analyses, PCA, Hierarchical clustering, K-Mean, SVM, Inductive Logic Programming, and Self Organizing Map
Language and methods**:** Prolog, Java/J2EE, Python, Perl, C++, PHP, R/BioConductor, GeneCluster, JTreeView,Weka, Prolog/ALEPH, IBM Cognos
Large Database design and implementation**:** IBM DB2, Mysql, Postgresql, MongoDB
Web application development**:** Tomcat, IBM WAS, JBOSS, SOAP, REST, Ajax, JSP/Servlet.
Software Architecture**:** UML, String Framework, JBOSS jBPM, IBM WebSphere
Distributed computing and Big Data: IBM InforSphere, Hadoop/MapReduce, HBase, MongoDB
Ability to multitask with strong organization/management, planning and problem solving skill.

Publications in computational biology

• Peer-reviewed journal articles

Nguyen H, Laurent M, Thompson JD, Poch O (2014). *Heterogeneous Biological Data Integration with High Level Query Language*. IBM Journal of Research and Development, vol. 58 no. 2/3 , 15 April, 2014. doi: 10.1147/JRD.2014.2309032 .

Bermejo-Das-Neves C., **Nguyen H**, Poch O and Thompson JD. (2014) *A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i)*, BMC Bioinformatics 2014, 15:111 doi:10.1186/1471-2105-15-111.(co-first author)

Nguyen H, Luu TD, Poch O, Thompson JD. (2013) *Knowledge Discovery from a Variant Database using Inductive Logic Programming.* Bioinformatics and Biology Insights.

Luu TD, Rusu AM, Walter V, Linard B, Poidevin L, Ripp R, Moulinier L, Muller J, Raffelsberger W, Wicker N, Lecompte O, Thompson JD, Poch O, **Nguyen H**. (2012). *KD4v: Comprehensible Knowledge Discovery System For Missense Variant*. Nucleic Acids Res; W71-75. (Corresponding – author: Hoan Nguyen)

Luu TD, Rusu AM, Walter V, Ripp R, Moulinier L, Muller J, Toursel T, Thompson JD, Poch O, **Nguyen H**. (2012). *MSV3d: database of human MisSense variants mapped to 3D protein structure.* Database (Oxford); bas018. (Corresponding – author: Hoan Nguyen)

Linard B, **Nguyen H**, Prosdocimi F, Poch O, Thompson JD (2012*). EvoluCode: evolutionary barcodes as a unifying framework for multilevel evolutionary data*. Evol. Bioinform Online;

Zeitz C, Jacobson SG, Hamel CP, Bujakowska K, Orhan E, Zanlonghi X, Lancelot ME, Michiels C, Schwartz SB, Bocquet B, CSNB consortium, Antonio A, Audier C, Letexier M, Saraiva JP, Luu TD, Sennlaub F, **Nguyen H**, Poch O, Dollfus H, Lecompte O, Kohl S, Sahel JA, Bhattacharya SS, Audo I. (2013) *Whole exome sequencing identifies mutations in LRIT3 as a cause for autosomal recessive complete congenital stationary night blindness*. Am J Hum Genet;

Audo I, Bujakowska K, *et al*. (2012). *Whole-exome sequencing identifies mutations in GPR179 leading to autosomal-recessive complete congenital stationary night blindness*. Am J Hum Genet; 90: 321-330.

Nguyen H, Wicker N, Kieffer D, Poch O. (2010) *A new projection method for biological semantic map generation*. J. Biomedical Science and Engineering; 3:13-19.

Friedrich A, Garnier N, Gagnière N, **Nguyen H**, Albou LP, Biancalana V, Bettler E, Deléage G, Lecompte O, Muller J, Moras D, Mandel JT, Toursel T, Moulinier L, Poch O. (2009) *SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases*. Hum Mutat. 31: 127-135

Audo I, Bujakowska K, Orhan E, Sennlaub F, Guillonneau X, Antonio A, Michiels C, Lancelot ME, Letexier M, Saraiva JP;**Nguyen H**, Luu TD, Léveillard T, Poch O, Paques M, Saddek MS, Bhattacharya S, Sahel JA, Zeitz C. (2013) *The familial dementia gene revisited: whole exome sequencing identifies a missense mutation in ITM2B underlying a novel autosomal dominant retinal dystrophy in a large family,* HMG.

Bard N, Bolze R, Caron E, Desprez F, Heymann M, Friedrich A, Moulinier L, **Nguyen NH**, Poch O, Toursel T (2010): *Decrypthon grid - grid resources dedicated to neuromuscular disorders*. Stud Health Technol Inform 2010, 159:124-133.

• Conference proceedings

Benabderrahmane S, Devignes MD, Malika Smail-Tabbone, Olivier Poch, Amedeo Napoli, Wolfgang Raffelsberger, Dominique Guenot, Nguyen Hoan, Eric Guerin. (2011) *Benchmarking a new semantic similarity measure using fuzzy clustering and reference sets: Application to cancer expression data*, 11th International Francophone Conference on Knowledge Extraction and Management EGC - 2011, Brest, France.

Luu TD, Nguyen N, Friedrich A, Muller J, Moulinier L, Poch O (2011). *Extracting knowledge from a mutation database related to human monogenic disease using inductive logic programming*, 2011 International Conference on Bioscience, Biochemistry and Bioinformatics IPCBEE vol.5 (2011) © (2011) IACSIT Press, Singapore.

- **Book Chapter**

Hoan Nguyen, Julie D.Thompson, Patrick Schutz and Olivier Poch. *Intelligent Integrative knowledge bases: bridging genomics, integrative biology and translational medicine*. In: Andreas Holzinger and Igor Jurisca. Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics. Springer LNCS, Volume 8401

Publications in computational astrophysics

- **Peer-reviewed journal articles**

Michel L, Motch C, Nguyen NH, Pineau FX (2014). *Building an archive with Saada. Astronomy and Computing*. Special Issue on The Virtual Observatory, Volumes 7–8, 2014, Pages 45–51.

- **Conference proceedings**

Michel L., Motch C., Pineau FX., **Nguyen H**. (2010) *Building Astronomical Databases with Saada (Update 2010)*. Astronomical Data Analysis Software and Systems XIX. Proceedings of a conference held October 4-8, 2009 in Sapporo, Japan. Edited by Yoshihiko Mizumoto, Koh-Ichiro Morita, and Masatoshi Ohishi. ASP Conference Series, Vol. 434. San Francisco: Astronomical Society of the Pacific, 2010. p.49. http://adsabs.harvard.edu/abs/2010ASPC..434..491M

Michel L., Motch C., **Nguyen H.**, Pineau FX. (2009) *A Guided Tour of Saada.* Astronomical Data Analysis Software and Systems XVIII ASP Conference Series, Vol. 411, proceedings of the conference held 2-5 November 2008 at Hotel Loews Le Concorde, Québec City, QC, Canada. Edited by David A. Bohlender, Daniel Durand, and Patrick Dowler. San Francisco: Astronomical Society of the Pacific, 2009. p.563. http://adsabs.harvard.edu/abs/2009ASPC..411..563M

Nguyen H., Michel L., Motch C. (2006) *Building an Astronomical Database with Saada*, Astronomical Data Analysis Software and Systems XV ASP Conference Series, Vol. 351, Proceedings of the Conference Held 2-5 October 2005 in San Lorenzo de El Escorial, Spain. Edited by Carlos Gabriel, Christophe Arviset, Daniel Ponz, and Enrique Solano. San Francisco: Astronomical Society of the Pacific, 2006. p.15. http://adsabs.harvard.edu/abs/2006ASPC..351...15N

Michel L., **Nguyen H.**, Motch C. *(2006) How to Publish Local Data Into the VO with Saada,* Astronomical Data Analysis Software and Systems XV ASP Conference Series, Vol. 351, Proceedings of the Conference Held 2-5 October 2005 in San Lorenzo de El Escorial, Spain. Edited by Carlos Gabriel, Christophe Arviset, Daniel Ponz, and Enrique Solano. San Francisco: Astronomical Society of the Pacific, 2006., p.25. http://adsabs.harvard.edu/abs/2006ASPC..351...25M

Michel L., **Nguyen H.**, Motch C. (2005) *SAADA: Astronomical Databases Made Easier.* Astronomical Data Analysis Software and Systems XIV ASP Conference Series, Vol. 347, Proceedings of the Conference held 24-27 October, 2004 in Pasadena, California, USA. Edited by P. Shopbell, M. Britton, and R. Ebert. San Francisco: Astronomical Society of the Pacific, 2005. p.71. http://adsabs.harvard.edu/abs/2005ASPC..347...71M

Nguyen H., Michel L., Motch C. (2004) *SAADA: An Automatic Archival System for Astronomy Data*. Astronomical Data Analysis Software and Systems (ADASS) XIII, Proceedings of the conference held 12-15 October, 2003 in Strasbourg, France. Edited by Francois Ochsenbein, Mark G. Allen and Daniel Egret. ASP Conference Proceedings, Vol. 314. San Francisco: Astronomical Society of the Pacific, 2004, p.121. http://adsabs.harvard.edu/abs/2004ASPC..314..121N